COMPUTER-ENHANCED KNOWLEDGE DISCOVERY IN ENVIRONMENTAL SCIENCE

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Doctor of Philosophy in
Environmental Science
at the University of Canterbury
by
Kyoko Fukuda

Abstract

Encouraging the use of computer algorithms by developing new algorithms and introducing uncommonly known algorithms for use on environmental science problems is a significant contribution, as it provides knowledge discovery tools to extract new aspects of results and draw new insights, additional to those from general statistical methods. Conducting analysis with appropriately chosen methods, in terms of quality of performance and results, computation time, flexibility and applicability to data of various natures, will help decision making in the policy development and management process for environmental studies. This thesis has three fundamental aims and motivations. Firstly, to develop a flexibly applicable attribute selection method, Tree Node Selection (TNS), and a decision tree assessment tool, Tree Node Selection for assessing decision tree structure (TNS-A), both of which use decision trees pre-generated by the widely used C4.5 decision tree algorithm as their information source, to identify important attributes from data. TNS helps the cost effective and efficient data collection and policy making process by selecting fewer, but important, attributes, and TNS-A provides a tool to assess the decision tree structure to extract information on the relationship of attributes and decisions. Secondly, to introduce the use of new, theoretical or unknown computer algorithms, such as the *K*-Maximum Subarray Algorithm (*K*-MSA) and Ant-Miner, by adjusting and maximizing their applicability and practicality to assess environmental science problems to bring new insights. Additionally, the unique advanced statistical and mathematical method, Singular Spectrum Analysis (SSA), is demonstrated as a data pre-processing method to help improve C4.5 results on noisy measurements. Thirdly, to promote, encourage and motivate environmental scientists to use ideas and methods developed in this thesis. The methods were tested with benchmark data and various real environmental science problems: sea container contamination, the Weed Risk Assessment model and weed spatial analysis for New Zealand Biosecurity, air pollution, climate and health, and defoliation imagery. The outcome of this thesis will be to introduce the concept and technique of data mining, a process of knowledge discovery from databases, to environmental science researchers in New Zealand and overseas by collaborating on future research to achieve, together with future policy and management, to maintain and sustain a healthy environment to live in.

Acknowledgements

Words are not good enough to express my gratitude as "*thank you*".

This thesis was completed because of the many people around me who supported and encouraged me to keep believing in myself. I couldn't have achieved this without all of you.

To my supervisors, Assoc. Prof. Jennifer Brown (Dept. of Math and Stats), Prof. Tadao Takaoka and Dr. Brent Martin (Dept. of Computer Science and Software Engineering), for providing me enormous support and understanding towards my thesis by commenting, editing and advising on my thesis as well as providing departmental financial support.

To all my family in Japan, who understood and allowed me to pursue my wish to study more.

To my partner, Phillip, who supported me by doing all our housework including looking after our many pets and me, when I was too busy concentrating on my work, who became my good adviser and editor for my research, who taught me programming, and gave me enormous support and programming help towards this thesis.

To my family in New Zealand, our corgis, Thumper and Minnie, and birds and fish, who always gave me smiles, and patiently waited for me to complete my thesis, so that they could play with me again.

I thank the departments and their administrative staff, who provided and helped me with grants, data, advice, and support. This thesis was supported by a University of Canterbury Doctoral Scholarship (2005-2008). The Department of Mathematics and Statistics accommodated my study and provided resources, conference funding and additional school fee finance for six months. The Department of Computer Science and Software Engineering provided a study scholarship in 2007.

I thank various people who showed an interest in my research and allowed me to investigate their data; Teresa Aberkane (ECan), Dr. Adrian McDonald and Dr. Andreas Baumgaertner (UC Physics), Dr. Carolyn Whyte and New Zealand Biosecurity team (MAF), Dr. Hamish Cochrane (UC Forestry), Dr. Peter Williams (Landcare), Dr. John Kean (AgResearch), Dr. Tomoko Nishida (NIAES), Dr. Mike Wulder (British Columbia), Dr. Ralf Wieland (ZALF), Dr. Gert Berger (ZALF), Prof. Andrew Hornblow (UC Health Sciences) and Prof. Ian Town (UC Deputy Vice Chancellor).

I also thank Prof. Town for bringing me the team for a future collaboration on an air pollution, climate and health study, and Dr. Phil Hider (Otago School of Medicine) for health research.

I would never forget to appreciate those people who believed in me; Prof. David Gunby (former Dean of Postgraduate Studies), Tracey Robinson (College of Science), Sue Clark (UC HR) and Dr. Jack Fergusson (UC Chemistry).

Thanks to Mr. Paul Brouwers and Mr. Steve Gourdie (UC Math) for excellent computer equipment support and their smiles!

To my examiners, Prof. Yasuharu Ukai (Kansai University) and Assoc. Prof. Earl Bardsley (Waikato University), who raised questions to contribute to improving my thesis.

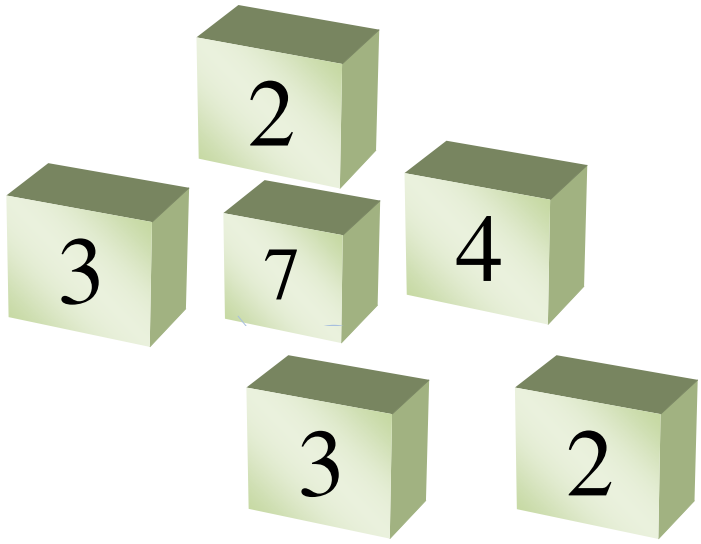
I couldn't have done this without all of you. Thank you!

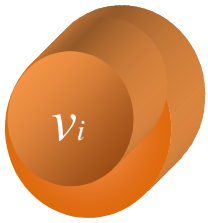
Table of Contents

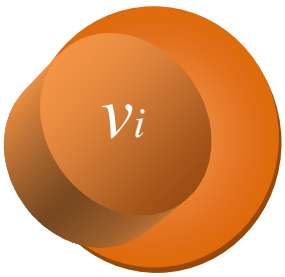
Table of Tables

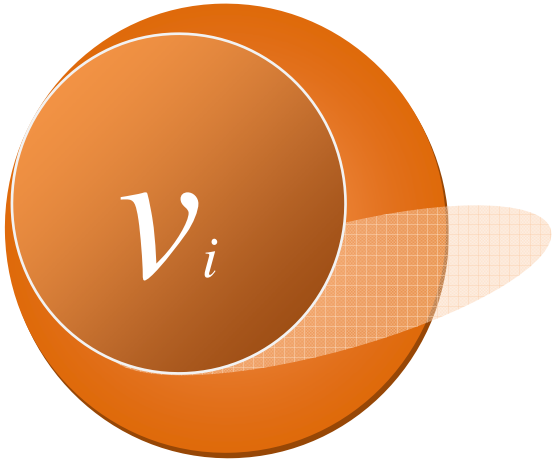
Table of Figures

Table of Appendices









Chapter 1. Data mining for environmental science problems

The motivation of this thesis is to enhance the use of commonly known computer algorithms and introduce little known computer algorithms as knowledge discovery concepts and tools, by adjusting, developing and validating them for future use in modelling for environmental science problems. This chapter introduces the concept of Knowledge Discovery in Databases (KDD) in environmental science by discussing how a KDD process, data mining, is applied to help in understanding of various environmental science problems. Each chapter in this thesis has its own theme, presenting new data mining techniques, the Tree Node Selection (TNS) method and Tree Node Selection for assessing decision tree structure (TNS-A), introducing computer algorithms, C4.5, *k*-means clustering, Ant-Miner, and *K*-Maximum Subarray (*K*-MSA), Singular Spectrum Analysis (SSA) as a data pre-processing method for C4.5, or demonstrating the use of data mining in various environmental science problems: New Zealand Biosecurity issues (Weed Risk Assessment model and sea container contamination), air pollution, climate and health, spatial weed distribution and defoliation. The outcome of this thesis will be to introduce such data mining tools and the KDD concept to various environmental science research organizations in New Zealand and overseas to help with future policy making and management processes, so that we can maintain and sustain a healthy environment to live in.

1.1. Knowledge discovery in environmental science

The motivation of this thesis is to enhance the use of commonly known computer algorithms and introduce uncommonly known computer algorithms as knowledge discovery concepts and tools, by adjusting, developing and validating them for future use in modelling environmental science problems.

Data mining is a process of knowledge discovery in databases (KDD), which involves multi-disciplinary fields such as machine learning, computer science, statistics, and pattern recognition. Data mining tools are used for prediction, e.g., classification, regression and time series, and knowledge discovery, e.g., deviation detection, database segmentation, clustering, and association rules (Weiss and Indurkhya 1998), shown in Fig. 1-1. The concept of data mining is well explained in various textbooks such as Mitchell (1997), which thoroughly describes the wide and deep concepts and algorithms of machine learning. Various newer data mining techniques in bioinformatics are introduced in Hsu (2006), statistical data mining and its comparison to computational data mining techniques are discussed in Hastie et al. (2001), and the data mining software WEKA is introduced in Witten and Frank (2005). Hence, this chapter focuses on the concept of how KDD can help understanding environmental science problems.

The term KDD has been defined as the "non-trivial extraction of implicit, previously unknown, and potentially useful information from data" (Frayley et al. 1991). This was revised to state that "KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al. 1996a,b). The concept

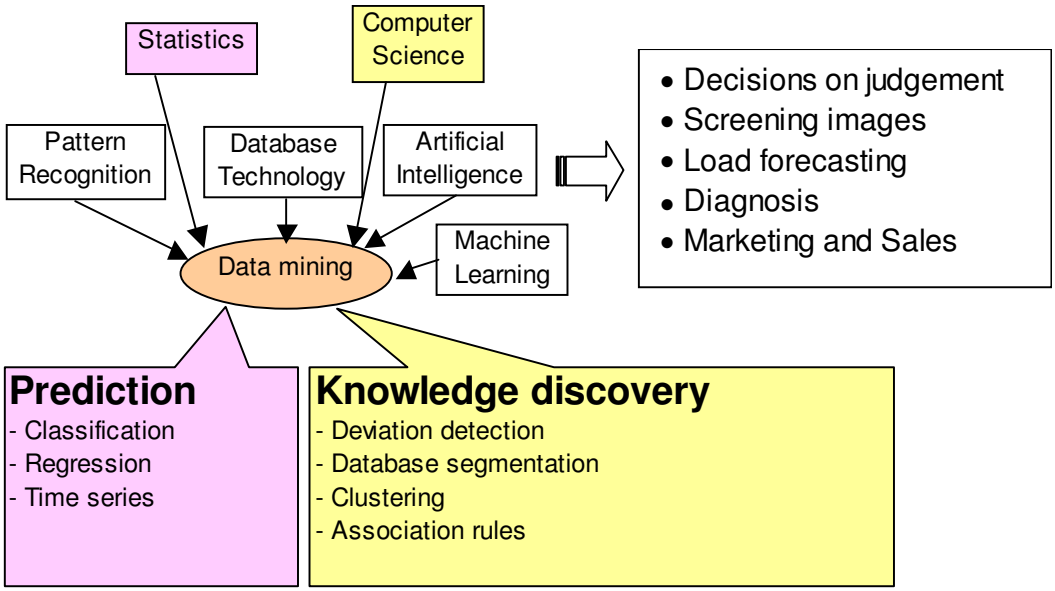


Fig. 1-1 Genesis of data mining with some examples (adapted from Weiss and Indurkhya 1998, Press 2004).

of KDD was introduced in 1989 at the first KDD workshop; the problem of extracting knowledge from data (or observations) is not new, but automation of processing large data sets opens up many new unsolved problems (Fayyad et al. 1996b). There have been arguments as to how KDD is different from other fields since the term data mining (computer based data analysis techniques) was first introduced in the 1960s. For example, statistics has a similar goal to KDD in *inference of knowledge from data* by providing a language and framework for quantifying the uncertainty resulting when one tries to infer general patterns from a particular sample of an overall population (details in Elder and Pregibon 1996; Fayyad et al. 1996b). Fayyad et al. (1996b) commented that concern arose over the fact that analysing large quantities of data (in any dataset, even randomly generated data) found patterns that appear to be statistically significant but in fact are not. There has been substantial progress in understanding such issues in statistics over the years, but data mining is a legitimate activity as long as one understands how to do it correctly. Furthermore, they stated that "*KDD can also be viewed as encompassing a broader view of modelling than statistics, aiming to provide tools to automate (to the degree possible) the entire process of data analysis, including the statistician's art of hypothesis selection"*.

1.2. Current use of KDD for environmental science applications

Since its introduction, KDD has played an important role in various sciences, e.g., web and text mining, social networking systems, bioinformatics, and business applications (KDD 2008). As one commonly-cited data mining is customer relationship management to develop business strategy. Ngai et al. (2009) investigated which data mining algorithms are commonly used in this field by surveying academic publications between 2000 and 2006, over 24 journals. They found that the most popular data mining techniques are, in order, neural networks, decision trees and association rules. In comparison to this, Spate et al. (2006), later published as Gibert et al. (2008), stated that artificial neural networks have also been applied extensively in the environmental sciences, but other data mining techniques were not found to be applied over a wide scale in the environmental sciences.

Artificial Neural Networks (ANNs) are generally applied in environmental science as a prediction tool. This could be because ANNs are flexible enough to be applicable to the nonlinear relationships and non-normality that often appear in environmental science data. Recent applications of ANNs are, for example, to predict stormwater quality at urbanized catchments located throughout the United States (May and Sivakumar 2009) and soil quality for forest data (Ito et al. 2008), and Wieland et al. (2006) developed a neural network tool box as a part of GIS software, particularly for environmental science problems. In other data

mining techniques, Walsh et al. (2008) and Liu et al. (2007) used clustering algorithms to interpret interannual ozone trends and to understand the spatial pattern of dead oak trees. A common use of decision tree algorithms is classification in environmental science applications; they are known to be used for land or feature classification in imagery. For example, Ozdogan and Gutman (2008) used them to process gridded climate and agricultural data as an advanced image classification algorithm, Tooke et al. (2009) used them to extract several urban vegetation characteristics, Goodwin et al. (2008) investigated the defoliation regions of the mountain pine beetle, and Elisabeth et al. (2006) used them to generate models for soil properties as a knowledge discovery tool.

Further data mining application examples were intensively reviewed by Spate et al. (2006) and Gibert et al. (2008) by introducing case studies, particularly regarding water quality or hydrology, e.g., Sànchez-Marrè et al. (1997), Comas et al. (2001) and Ter Braak et al. (2003). Additionally, each chapter in this thesis contains its own literature review for methodology related to the chapter topic.

1.3. Motivations for the use of data mining techniques for environmental science problems

Data mining algorithms are often flexible to various natures of data, e.g., missing and incomplete, Boolean, continuous, discrete, categorical and textual. They are generally designed to be applied on very large data sets, e.g., millions of data points for text recognition. Larger data sets are considered to better represent reality, which allows better rules to be generated and more reliable results to be obtained. Statistical analyses are comparable to data mining approaches; for example, logistic regression is often compared with the well known C4.5 algorithm (Quinlan 1993) as both are off-the-shelf methods for building classification models (Perlich et al. 2003). However, such statistical methods tend to be computationally expensive, thus, only small to medium data generally suit thorough statistical investigation (will be discussed in Chapter 3, Study II).

Environmental science data sets can be large, e.g., temperature measurements could be collected over hundreds of years, but they often consist of small to medium amounts of data, e.g., a few hundred to several thousand instances with less than twenty attributes, in particular for ecological data, as the data collection process can be limited by time, labour and equipment cost, availability of enough relevant data, ill structured data due to equipment errors or maintenance, or the recency of the problem. For example, the New Zealand Biosecurity Act was enacted in 1993 to provide for the effective management of risks associated with the importation of risky goods (Ministry of Agriculture 1993). The first New

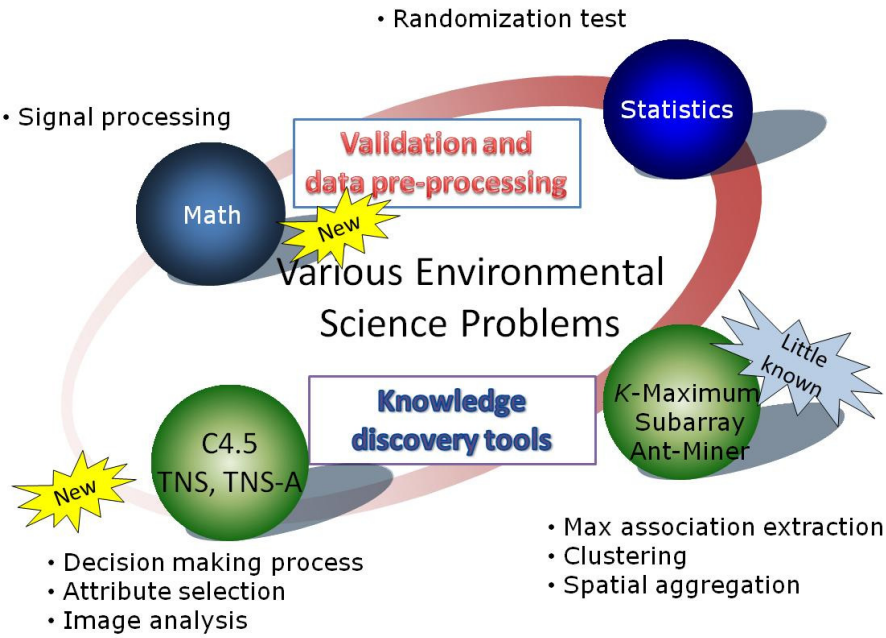


Fig. 1-2 Framework of this thesis.

Zealand Biosecurity Strategy was developed in 2003 (discussed in Chapter 3, Study II). A greater awareness of biosecurity issues was raised since 2001 in response to events such as the outbreak of foot and mouth disease in the United Kingdom (MAF 2002). Hence, the collection of biosecurity data and the search for appropriate quantitative analysis tools to investigate the problem are still in progress.

For this point, it is important that the analysis can be conducted with an appropriate choice of methods, in terms of quality of performance and results, computation time, flexibility and applicability to data of various natures, so that results of environmental studies can help decision making in the policy development and management process. At the same time, it is critical to the efficiency and effectiveness of the data collection process to select fewer but higher quality variables, attributes, predictors or questions that describe the problem, so that the data collection strategy is successful to minimise the cost of future collection or analysis. Both data mining and statistical approaches can be applied on large as well as small data sets, and it would be simple to conclude that such small data sets may not be worthwhile to investigate as the data structures are not yet good enough to find statistically significant results. However, in reality, the availability of large or complete data sets is limited; it is ideal to find a flexible method that extracts some knowledge from data of various natures, and statistical methods can be used to help validating, supporting or improving the data mining results.

This thesis introduces the combined applications of computer algorithms as knowledge discovery tools, and mathematical and statistical methods as validation and data pre-processing for data mining techniques. Fig. 1-2 demonstrates the framework of various methods that were proposed in this thesis and will be introduced in the next section.

The overall goal of this thesis is to introduce how various environmental science problems can be investigated with computer algorithms and how the knowledge extracted from data helps understanding problems for the future policy making and management process to improve and maintain the health of our environment. At the same time, this thesis introduces how mathematical and statistical methods can validate and assist both inputs and outputs of data mining techniques.

1.4. How to read this thesis

The rest of the thesis consists of five chapters (Chapters 2-6), each of which has its own motivation, and consists of abstract, introduction, methods, results and discussion, and conclusions sections. The final chapter, Chapter 7, describes future plans resulting from this thesis. Below is a brief description of each chapter, describing my tasks and contributions to this thesis.

- **Chapter 2. Introducing a new attribute selection method: Tree Node Selection**

It would be valuable if we could minimise the amount of data to collect but maintain the quality of the results and the performance of the algorithm. This chapter covers the general concept of attribute selection (AS) methods and the development of a new ranking filter attribute selection tool, the Tree Node Selection (TNS) method. The TNS selects a smaller, but more relevant set of attributes using decision trees pre-generated by the well known C4.5 algorithm (Quinlan 1993) as its information source. The performance of TNS was compared with five well known AS methods on 33 benchmark data sets (UCI database, Asuncion and Newman 2007) using the C4.5 (pruned and unpruned) and naïve Bayes classifiers. Results were assessed using a combination of various statistical methods, e.g., ANOVA, the Kruskal-Wallis hypothesis test, and Tukey's test for multiple comparisons. The motivation behind the development of TNS was to bring attribute selection closer to the decision making process by directly analysing the decision tree structure. The part of this work was presented in Fukuda and Martin (in press).

My contribution in this chapter was to develop the idea, concept and detail of the TNS algorithm, and evaluate and validate its use. Dr. Martin advised me on the general concepts of

attribute selection and data mining, Prof. Takaoka reviewed the TNS algorithm, and Assoc. Prof. Brown advised me on the statistical validation process.

- **Chapter 3. Application of TNS and TNS-A for environmental science studies**

Firstly, this chapter introduces the application of TNS and the Ant-Miner algorithm (Parpinelli et al. 2002) on the Weed Risk Assessment (WRA) model (Pheloung et al. 1998). Secondly, this chapter introduces the development of a new assessment tool for decision tree structure, Tree Node Selection for assessing decision tree structure (TNS-A), and the application of TNS and TNS-A on the risk profiles of the sea container contamination pathway. Both of these case studies are involved in the governmental decision making process in biosecurity, for preventing the entry of unwanted or contaminated alien plants or containers into the country to protect the health of the environment in New Zealand. This chapter compares TNS and Ant-Miner as attribute selection methods. The application of TNS and development of TNS-A aimed to extract information on the relationships between pairs of attributes and between attributes and decisions (classes) in decision tree structures, to help understanding the decision making process by identifying what questions and items were important to improve the biosecurity strategy. This chapter briefly discusses the different approaches that are taken to analyse the unique nature of the sea container data, i.e., all text with many unique variables, between data mining and statistical methods. The application of TNS and Ant-Miner on the WRA model was presented in Fukuda and Brown (2007a,b) and the sea container contamination analysis was presented to advise the Ministry of Agriculture and Forestry (MAF), New Zealand, for the future use of data mining techniques.

The application and interpretation of Ant-Miner, the idea and development of the TNS-A algorithm, interpretation and discussion of results were my contributions. Prof. Takaoka reviewed the TNS algorithm, Assoc. Prof. Brown helped with the WRA findings, and Dr. Whyte (MAF) provided current knowledge of the sea container contamination problems and reviewed and edited findings in this chapter.

- **Chapter 4. Introducing the *K*-Maximum Subarray Algorithm (*K*-MSA) for studying air pollution, climate and health**

This chapter covers the basic concept of the *K*-MSA (Bae and Takaoka 2006; 2007) and demonstrates the practical use of the *K*-MSA as a knowledge discovery tool for an air pollution, climate and health study. Generally, air pollution, climate and health studies are investigated by time series approaches (Fukuda 2004; Fukuda and Hudson 2005a,b), whereas the application of computer algorithms is not yet common. The *K*-MSA investigates the air pollution level and the age of the admitted patient, and various climate variables and the age

of the admitted patient, by forming a two-dimensional array and locating maximum subarrays containing clusters of high acute respiratory admission rates in Christchurch. The preliminary work was presented in Fukuda and Takaoka (2007a).

The *K*-MSA investigation and interpretation of air pollution, climate and health contents and results were my contribution. Prof. Takaoka advised on the algorithm of the *K*-MSA, and Dr. Hider (School of Medicine and Health Sciences, University of Otago) reviewed and edited health findings.

- **Chapter 5. Exploring the *K*-MSA as an alternative to clustering for environmental science data**

As the *K*-MSA detects the maximum aggregated data points over a two-dimensional array, it can act like a clustering method. This chapter introduces the new use of *K*-MSA in two clustering problems involved in environmental science studies: the Bumpus sparrow (benchmark data) and the spatial weed aggregation patterns. The main development in this chapter was to adjust and improve the *K*-MSA applicability and practicality for the environmental science investigation. This chapter covers the concept of a new parameter, the weight parameter, which was developed to increase detection sensitivities for *K*-MSA maximum subarray regions, presented in Fukuda and Takaoka (2007b). This chapter also introduced the new concept of the randomisation test to assess the ecological significance of the *K*-MSA results. In order to validate the *K*-MSA as an alternative to clustering methods, this chapter compared the results between *k*-means clustering and the *K*-MSA, and between the ecological clustering method, Spatial Analysis by Distance IndicEs (Perry 1995; Perry et al. 1999) and the *K*-MSA. The spatial weed aggregation study was presented in Fukuda et al. (2008).

Comparison of the *k*-means clustering method, introducing the concept and development of a new weight parameter for the *K*-MSA, and investigation and interpretation of the spatial weed aggregation patterns via the *K*-MSA were my contributions. Assoc. Prof. Brown brought the idea to incorporate the randomisation test and tested the SADIE method. Dr. Williams (Ecologist at Landcare Research, Nelson) and Dr. Kean (AgResearch, Lincoln) provided knowledge about the weed data, and Dr. Williams reviewed the findings.

- **Chapter 6. Singular Spectrum Analysis for decision tree classifier**

This chapter introduces two experimental environmental science case studies involving noisy inputs for decision tree construction: improved decision tree construction for climate and air pollution, and a comparison of data mining and image segmentation approaches for classifying defoliation regions in aerial forest imagery. Environmental science data, air

pollution and climate measurements in particular, are known to be noisy, which can confuse algorithms and cause poor prediction results. This chapter covers the exploratory use of the unique mathematical decomposition method, Singular Spectrum Analysis (Golyandina et al. 2001), as a pre-processing method for noisy input data to improve data mining classification for C4.5. The investigations were presented in Fukuda (2007) and Fukuda and Pearson (2007a,b).

The concept, ideas and implementation of SSA as the data pre-processing method for the decision tree application for both studies were my contributions. The developed approaches that were taken for image analysis using SSA were my contributions. The co-author of Fukuda and Pearson (2006a,b), Mr. Pearson, developed the simple clustering algorithm to compare with the data mining approach in the defoliation imagery study.

- **Chapter 7. Conclusions: Future plans and software development for environmental science problems.**

The outcome of this thesis is to introduce and encourage attribute selection methods, newly developed computer algorithms, and commonly and uncommonly known computer algorithms to various environmental studies and scientists to help with their decision making systems by improving analysis from discovering new aspects from data, in addition to general statistical analyses. This chapter discusses future plans and collaboration projects that arose from this thesis. From the knowledge that I gained from this thesis and in partnership with other researchers, I applied for and obtained two research grants; firstly, funding from the Royal Society of New Zealand ISAT Linkages Fund (obtained August 2008) for *K*-MSA GIS software development for ecological modelling, and secondly, a General Project Grant from the Canterbury Medical Research Foundation (obtained September 2008) to develop a hybrid prediction model using data mining, computer algorithms and statistics for air pollution, climate and health. This chapter covers future research plans with the National Institute for Agro-Environmental Sciences in Japan for the weed analysis, the Leibniz Centre for Agricultural Landscape Research in Germany for the GIS model, and the School of Medicine and Health Sciences, University of Otago, for air pollution, climate and health research. I will be working for these projects as the principal investigator to keep challenging and continuing to bridge between various environmental sciences problems and scientists, in order to maintain and improve the environment in which we live.

1.5. References

Asuncion A, Newman DJ (2007) UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science. Available via http://mlearn.ics.uci.edu/MLRepository.html. Accessed on 28 July, 2008.

Bae SE, Takaoka T (2006) Improved algorithms for the *K*-Maximum Subarray problem. Comput J 49:358-374.

Bae SE, Takaoka T (2007) Algorithms for *K*-Disjoint Maximum Subarrays. Int J Found Comput Sci 18:319-339.

Comas J, Dzeroski S, Gibert K (2001) Knowledge discovery by means of inductive methods in wastewater treatment plant data AI Com 14:45–62.

Elder J, Pregibon DA (1996) A statistical perspective on KDD. In: Fayyad U, Piatetsky-Shapiro G, Smyth P (eds). In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Cambridge, Mass.

Elisabeth NB, Henderson BL, Viergever K (2006) Knowledge discovery from models of soil properties developed through data mining. Ecol Mod 191:431-446.

Fayyad U, Platetsky-Shapiro G, Smyth P (1996a) From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in Knowledge discovery and data mining. MA/MIT Press, Cambridge.

Fayyad U, Piatetskay-Shapiro G, Smyth P (1996b) The KDD process for extracting useful knowledge from volumes of data, Commun ACM 39: 27 – 34.

Frayley W, Piatetsky-Syapiro G, Matheus C (1991) Knowledge discovery in databases: An overview. In: Piatetsky-Shapiro G, Frawley W (eds.) Knowledge Discovery in Databases, MA/MIT Press, Cambridge.

Fukuda K (2004) New improved methods for application and interpretation of SSA: A case study of climate and air pollution in Christchurch, New Zealand", MSc thesis, University of Canterbury, Christchurch, New Zealand.

Fukuda K (2007) Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution. In Proc. of the 2007 IEEE CIDM 2007, 697-704.

Fukuda K, Brown J (2007a) Investigation of The Weed Risk Assessment Model Using Data Mining, Intl Conf. of 9th EMAPi9, abstract.

Fukuda K, Brown J (2007b) Classification Rule Extraction by Ant-Miner for Weed Risk Assessment, In Oxley L and Kulasiri D. (eds) MODSIM 07, 2882-2888.

Fukuda K, Brown J, Williams P, Kean J (2008) The *K*-Maximum Subarray algorithm as an alternative clustering analysis for the spatial weed aggregation pattern, NZSA 2008, abstract.

Fukuda K, Hudson IL (2005a) Global and local climatic factors on sulfur dioxide levels: comparison of residential and industrial sites, In Proc. of 20th IWSM, 187-194.

Fukuda K, Hudson IL (2005b) Investigations of short-term (hourly) weather influences on CO, NO, NO_2, PM_{10} and SO_2 Levels in Christchurch, New Zealand, In Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural System, 45-52.

Fukuda K, Martin B (in press) Decision Trees as Information Source for Attribute Selection, In Proc. of the 2009 IEEE CIDM, 0-8.

Fukuda K, Pearson PA (2006a) Investigation of Singular Spectrum Analysis and Machine Learning for Road Sign Location. In Extended Abstracts of 7th DAS 2006, 29-32.

Fukuda K, Pearson PA (2006b) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-6.

Fukuda K, Takaoka T (2007a) Analysis of Air Pollution (PM_{10}) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm, In Proc. of the 2007 ACM SAC 2007, 153-157.

Fukuda K, Takaoka T (2007b) Investigation of the maximum association for suicide rate and social factors using Computer Algorithm, In Oxley L and Kulasiri D. (eds) MODSIM 07, 1381-1387.

Giberta K, Spate J, Sànchez-Marrèc M, Athanasiadisd IN, Comase J (2008) Chapter 12: data mining for environmental systems, Dev I Env Ass 3: 205-228.

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, Boca Raton.

Goodwin NR, Coops NC, Wulder MA, Gillanders S, Schroeder TA, Nelson T (2008) Estimation of insect infestation dynamics using a temporal sequence of Landsat data, Rem Sen Environ 112: 3680-3689.

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning, data mining, inference, and prediction. Spring

Ito E, Ono K, Ito Y-M, Araki M (2008) A neural network approach to simple prediction of soil nitrification potential: A case study in Japanese temperate forests. Ecol Mod 219: 200-211.

KDD (2008) Research track accepted papers, The 14th ACM SIGKDD, http://www.sigkdd.org/kdd2008/papers.htmls. Accessed on 26 January 2009.

Liu D, Kelly M, Gong P, Guo Q (2007) Characterizing spatial–temporal tree mortality patterns associated with a new forest disease, For Ecol Manag 253: 220-231.

MAF (2002) Report of the controller and auditor-general, Ministry of Agriculture and Forestry: Management of biosecurity risks, Wellington.

May DB, Sivakumar M (2009) Prediction of urban stormwater quality using artificial neural networks, Environ. Mod Softw 24: 296-302.

Ministry of Agriculture (1993) Biosecurity Act 1993. Available via http://www.legislation.govt.nz/act/public/1993/0095/latest/whole.html#DLM314623. Accessed on 26 September 2008.

Mitchell TM (1997) Machine learning, McGraw-Hill, New York.

Ozdogan M,Gutman G (2008) A new methodology to map irrigated areas using multi-temporal MODIS and ancillary data: An application example in the continental US, Rem Sen Environ 112:3520-3537.

Parpinelli RS, Lopes HS, Freitas AA (2002) Data Mining With and Ant Colony Optimization Algorithm, IEEE Trans Evol Comput 6: 321-332.

Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: a learning-curve analysis. JMLR 4: 211-255.

Perry JN (1995) Spatial Analysis by Distance Indices. J Anim Ecol 64: 303-314.

Perry JN, Winder L, Holland JM, Alston RD (1999) Red-blue plots for detecting clusters in count data. Ecol Let 2:106-113.

Pheloung PC, Williams PA, Halloy SR (1999) A weed risk assessment model for use as a biosecurity tool evaluating plant introductions, J Environ Manag 57: 239-251.

Press SJ (2004) The role of bayesian and frequentist multivariate modeling in statistical data mining. In: Bozdogan H (ed), Statistical Data Mining and Knowledge Discovery, Chapman & Hall/CRC, Boca Raton; 1-14.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Sànchez-Marrè M, Cort´es U, B´ejar J, de Gracia J, Lafuente J, Poch M (1997) Concept information in wastewater treatment plants by means of classification techniques: an applied study. Appl Int 7: 146-166.

Spate JM, Gibert K, Sànchez-Marrè M, Frank E, Comas J, Athanasiadis I, Letcher R (2006) Data mining as a tool for environmental scientists. In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-22, Burlington.

Ter Braak C, Hoijtink H, Akkermans W, Verdonschot P (2003) Bayesian model-based cluster analysis of predicting macrofaunal communities Ecol Mod 160: 235–248.

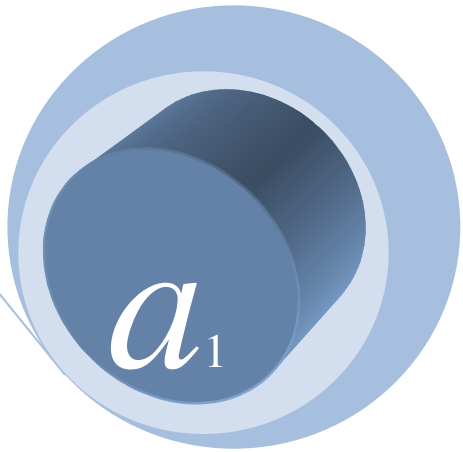
Tooke TR, Coops NC, Goodwin NR, Voogt JA (2009) Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications, Rem Sen Environ 113: 398-407.

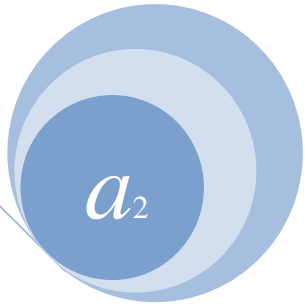
Walsh KJ, Milligan M, Woodman M, Sherwell J (2008) Data mining to characterize ozone behavior in Baltimore and Washington, DC, Atmos Environ 42: 4280-4292.

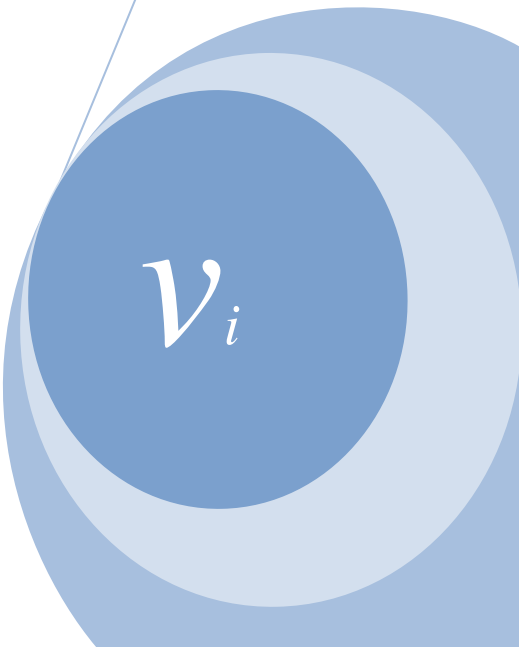
Weiss M, Indurkhya N (1998) Predictive Data Mining: A Practical Guide. Morgan Kaufmann, San Francisco.

Wieland R, Voss M, Holtmann X, Mirschel W, Ajibefun I (2006) Spatial analysis and modeling tool (SAMT): 1. Structure and possibilities. Ecol Inf 1:67-75.

Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2nd edn.







Chapter 2. Introducing a new attribute selection method: Tree Node Selection (Fukuda and Martin, in press)

Attribute selection (AS) is known to help improve the results of algorithmic learning processes by selecting the few input attributes that are the most predictive. This study introduces a new ranking filter AS method, the Tree Node Selection (TNS) method as a knowledge discovery tool. In the manner of a *pruning* process, it produces a concise tree and improves results by removing less relevant information from a pre-generated decision tree. TNS selects a smaller, but more relevant set of attributes by analyzing the existing decision tree, counting the number of instances that are classified by paths passing through each node or leaf node to assess the significance of each attribute. To test the performance of TNS, 33 benchmark datasets (UCI) with various numbers of instances, attributes and classes were investigated along with five known AS methods, and the results were tested with the C4.5 (pruned and unpruned) and naïve Bayes classifiers. The performance, in terms of classification accuracy improvement, reduction in the number of attributes and the size of the generated decision tree are assessed by various statistical analyses for multiple comparisons. Additionally, the performance differences between pruned and unpruned decision tree construction, and the processing time for all AS methods, were assessed separately. Overall results suggest that TNS is the most consistent AS method to simultaneously achieve good classification accuracy, number of attributes and decision tree size for all classifiers.

2.1. Introduction

The concept of attribute selection (AS) is particularly helpful for environmental science problems, because identifying the most relevant or predictive attributes for a given problem contributes to cost effective data collection and management. Many environmental science investigations are involved in governmental or regional decision making and management processes, therefore my goal in developing the Tree Node Selection method, TNS, was to assess the structure of decision trees to identify attributes that are important for decision making. TNS is an AS method, but additionally, TNS for assessing decision tree structure, TNS-A, was developed to assess the decision tree structure to extract relationships between attributes and decisions in the tree as a knowledge discovery tool.

This chapter covers the concept of AS in brief, details of the TNS algorithm and the benchmark experiment to assess the performance of TNS over various existing AS methods. Fukuda and Martin (in press) described part of the investigation in this chapter, from introducing how various attribute selection methods, in particular TNS, improved the unpruned decision tree construction using the same 33 benchmark data as this chapter.

2.1.1. Motivations of attribute selection in environmental science

Irrelevant, redundant or noisy features can confuse learning algorithms and cause them to construct poor classifiers (Last et al. 2001). A number of attribute selection techniques, e.g., Information Gain (Quinlan 1993) and Relief (Kira and Rendell 1992; Kononenko 1994), have been introduced and tested by researchers, e.g., Hall and Holmes (2003), to improve the results or clarity of classifiers by selecting the set of attributes or features that are most predictive of the outcome, without loss of the original meaning of the attributes after the reduction.

Attribute selection (AS) methods are generally structured as *wrapper* or *filter* approaches. The *wrapper* selects an attribute subset that is optimized for a given classification algorithm, treated as a black box, by repeatedly running the algorithm on many candidate subsets and measuring the quality of the subset each time. Thus, *wrapper* is generally known to provide better results due to the interaction between the search and the learning scheme's inductive bias (Hall and Holmes 2003), but it is a time consuming method that is not practically advantageous. The *filter* selects an independent attribute subset of the classification algorithm, and is less time consuming (Freitas 2002). Generally, AS is applied as a data pre-processing step to sufficiently reduce the number of attributes from the thousands that, for example, commonly occur in text and web classification problems, so that the computational

complexity can be minimized, but AS can also act as a knowledge discovery tool for small to medium sized data sets (Jensen and Shen 2007), so that irrelevant attributes can be identified to aid data collection and management.

Most environmental problems consist of unknown factors until the initial experiment provides some knowledge about the data. It is difficult to identify how and which attributes are important for the problem, besides there are limitations with respect to cost of experiments and accessibility of data. Also, many environmental science investigations are involved in governmental or regional decision making, to propose management processes and strategies to mitigate and control problems, in order to improve the health of the environment. Therefore, it is desirable to select a tool to assess and quantify the problems or increase knowledge about data. In regard to the attribute selection approach, traditional statistical analysis, e.g., principal components and regression, is commonly used in environmental science to reduce the dimension of input attributes, but can be computationally expensive (see detailed discussion in Chapter 3, Study II). Data mining techniques tend to be computationally efficient as they are generally designed to process large data sets, e.g., millions of instances for text mining problems, and are flexible in the nature of data; they can handle Boolean, continuous, discrete, and missing data points. However, environmental science data tends to consist of small to medium numbers of instances and attributes, e.g., thousands, rather than tens of thousands or millions of data points. Spate et al. (2006) commented that choice of data mining methods should also be influenced by data size. They advised the use of simple methods, such as the commonly known C4.5 algorithm, and to be mindful of the maximum theoretical certainty; as the number of data increases, variance of classical estimators tends to zero, which implies that small sample differences may appear statistically significant.

2.1.2. The C4.5 decision tree algorithm

Decision tree learning, e.g., C4.5 (Quinlan 1993), is practically and widely used as a simple classification data mining technique for inductive inference (Mitchell 1997) and describes the decision process in a readable, comprehensible manner. Previously, modifications to the AS stage of the C4.5 algorithm have been suggested, as follows. Wang and Jiang (2007) proposed an *average gain* measure to penalize attributes with many values by dividing the gain by the number of attribute values, to single out an improved attribute. This measure is an alternative to *gain ratio*, which penalizes attributes with many values by incorporating a term called *split information* (details in Quinlan 1986, 1987, 1993). Wang et al. (2002) constructed a decision tree with attributes that highly contribute to the decision tree

classification based on rough set theory (Pawlak 1991), measuring data dependencies by recognizing sensitive or insensitive attributes, instead of the entropy of information. However, this study uses C4.5 as the information source for the attribute selection method developed in this thesis, TNS.

2.1.3. Motivations of Tree Node Selection (TNS) method

In this study, I have developed the Tree Node Selection (TNS) method, a new data mining attribute selection method, to select the most predictive attributes for the model. The uniqueness of TNS is that it obtains the information it needs to select attributes by analyzing a pre-generated decision tree, in the same manner as a *pruning* method, considering each of the decision nodes in the tree to produces a concise tree, and improving results by removing irrelevant information. Generally, *rule post-pruning*, used for C4.5 (Quinlan 1993) generates a rule for each leaf node in the tree, with each attribute test along the path from the root to the leaf becoming a rule antecedent (precondition) and the classification at the leaf node becoming the rule consequent (postcondition). Then, for each rule, *rule post-pruning* removes any preconditions that result in improving its estimated accuracy, and finally sorts the rules by estimated accuracy (Mitchell 1997). In comparison, *reduced-error pruning* (Quinlan 1987) replaces subtrees with leaf nodes, assigning each the most common classification of the training examples affiliated with the node at the root of the subtree, where such pruning can be done without decreasing classification accuracy. TNS assesses each node in the tree by counting the number of instances that are classified by a path passing through the node. TNS then ranks the overall contribution for each attribute by the sum of such instance counts for all nodes labelled with the given attribute. The attribute set is repeatedly evaluated on the training data, removing the lowest ranked attribute each time. This process continues until the attribute set is empty, finally choosing the attribute set that performed best, i.e., had the highest classification accuracy obtained through the ranking filter method (details are described in Section 2.2.2). Thus, the selected attributes from TNS help understanding which factors (ranked attributes) are important for the decision support system. An interesting part of the TNS approach is that it can explore either a pruned or an unpruned decision tree to select predictive attributes, but regardless of whether pruning was used or not, the selected attributes can be used to generate a final decision tree either with or without pruning, to obtain the classification accuracy. This is due to the practical observation that not all trees obtain better classification accuracy after pruning, even though the purpose of pruning is to obtain a smaller tree and prevent over-fitting, to help improve the classification accuracy.

Another interesting approach of TNS is that the decision tree algorithm (a *greedy algorithm*) may select a root node that is not globally optimal; it cannot always be assumed that the selected root node uses the most important or influential attribute for the entire decision process. TNS identifies or ranks the attributes by counting the number of instances that are classified by paths including each node in the decision tree. TNS can rank attributes that appear often in the tree as more important than attributes that are closer to the root node, although it is very likely that the attribute used for the decision from the root node will be ranked by TNS as the most important attribute in the tree.

In this chapter, the two TNS methods, based on the pruned and unpruned output from the decision tree algorithm J4.8, in WEKA 3.4.11 (WEKA 2008), based on C4.5 (Quinlan 1993), are tested on 33 benchmark experiments from the UCI collection (Asuncion and Newman 2007). Five different freely available AS methods in WEKA, which were also tested by Hall and Holmes (2001), were selected for comparison with TNS: ranking attribute selectors Information Gain Attribute Ranking (IG) and Relief (RLF), subset evaluators Correlation-based Feature Selection (CFS) and Consistency-based Subset Evaluation (CNS), and a wrapper method (WRP). Note that the principal component analysis, which defines new attributes as linear combinations of existing attributes, as opposed to selecting a subset of attributes as like all other AS methods, is also available in WEKA. The classification accuracy (10-fold cross validation, n=10) using the selected attributes is evaluated by a decision tree algorithm (C4.5), which employs a top-down, greedy search through the space of possible decision trees, and the naïve Bayes classifier, a Bayesian learning method, which

Table 2-1 Description of 33 benchmark datasets (sorted by number of instances).

Data set	Instances	Attributes	Classes	Data set	Instances	Attributes	Classes
labor-relations	40	16	2	balance-scale	625	4	3
zoo	101	16	7	soybean	683	34	19
iris	150	4	3	credit-screening	687	15	2
hepatitis2	155	19	2	pimadiabetes	768	8	2
wine	178	13	3	vehicle	846	18	4
flags	194	26	8	anneal	898	38	5
sonar	208	60	2	mammographic	961	5	2
audiology	226	69	24	german credit	1000	20	2
breast-cancer2	286	9	2	splice	3190	61	3
horse-coli	300	22	2	kr-vs-kp	3196	36	2
heart-c	303	13	2	hypothyroid	3772	29	4
ecoli	336	7	8	segment	5000	19	7
primary-tumor	339	17	21	waveform	5000	21	3
ionophere	351	34	2	mushroom	8124	22	2
voting	435	16	2	letter recognition	20000	16	26
arrhythmia	452	279	13	adult	48841	14	2
bands	538	39	2				

calculates explicit probabilities for hypotheses (Mitchell 1997). Two different types of classifier algorithms were applied to test AS methods: C4.5 classifies instances by sorting them down the tree from the root to some leaf node, and naïve Bayes classifies instances to the class with highest probability (Mitchell 1997; Hall and Holmes 2003). The performance of the different AS methods is assessed by various statistical methods, e.g., Analysis of Variance (ANOVA), Tukey's and Kruskal-Wallis test, and lastly, the processing time for all AS methods is assessed.

2.2. Data and methods

2.2.1. Data preparation

The 33 benchmark data sets, with the number of instances ranging from forty to tens of thousands, attributes ranging from four to 279, and classes ranging from two to 26, shown in Table 2-1, were investigated for the performance of a total of seven AS methods: two types of Tree Node Selection (TNS), TNSP and TNSU, that take input information from pruned and unpruned decision trees, respectively, Information Gain Attribute Ranking (IG), Relief (RLF), Correlation-based Feature Selection (CFS), Consistency-based Subset Evaluation (CNS) and wrapper (WRP) based on C4.5 and naïve Bayes (NB).

2.2.2. Attribute selection process

All AS methods were assessed by comparing the mean and standard deviation ($\mu \pm SD$) of their classification accuracies (obtained using 10-fold cross-validation[1]) over the 33 data sets, described in detail as follows.

The AS process has three steps for each cross-validation fold: attribute selection (step 1), attribute evaluation (step 2) and attribute testing (step 3), shown in Fig. 2-1. Hence, ten-fold cross-validation ($n = 10$) was used for the pruned and unpruned C4.5 decision tree and NB classifiers, giving in total $n = 330$ for each method. Initial attribute selection (step 1 in Fig. 2-1) starts by splitting data into a training set (90%) to run a single attribute selection method to evaluate or select attributes for evaluation (step 2). During step 1, all data were discretised for IG, CFS and CNS, since these methods require discretised input data before AS. Selected attributes are then tested on a testing set (10%) using each of the pruned and unpruned C4.5

[1] The 10-fold cross-validation is a commonly used method for estimating the accuracy of a generated rule from a limited set of data and the final classification accuracy is a result of the 10-fold cross-validation. The dataset is divided into a number of partitions (e.g., 10), each of which is used as the test set for a decision tree generated using the remainder of the data. Each fold is generated fairly, by the same method used by the WEKA software (details in WEKA 2008).This allows all of the data to be used for both training and testing. However, note that the default setting of WEKA software (WEKA 2008) uses the entire data set to generate a tree for display, and tests it using 10-fold cross-validation.

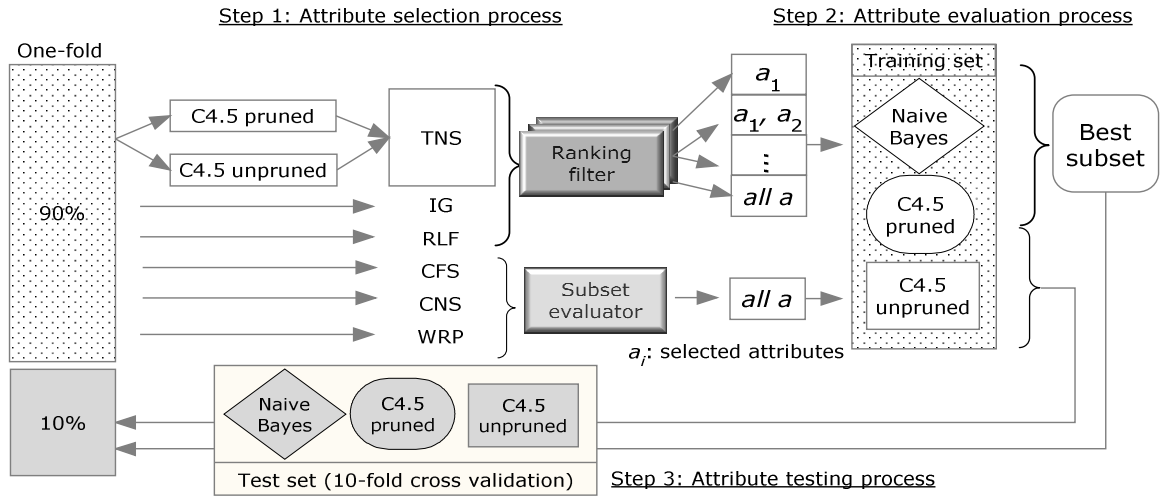


Fig. 2-1 Attribute selection process.

decision tree and NB classifiers to provide the per-fold classification accuracy, i.e., one classification accuracy for a single fold, for the attribute testing step (details in step 3, as follows). Note that during the attribute selection process (step 1), TNS evaluates the pruned and unpruned C4.5 decision trees that were generated from the training set.

Attribute evaluation (step 2 in Fig. 2-1) selects the final attribute set to be tested for each AS method. Subset evaluators such as CFS, CNS and WRP assess attributes from a training set and provide a subset of (selected) attributes, which is passed through directly to attribute testing. Ranking filter methods such as TNS, IG and RLF assess attributes from a training set and output them in ranked order, indicating, for example, that attributes A-F are selected, with A ranked highest and F ranked lowest. The attributes are then evaluated using the pruned and unpruned C4.5 decision tree and NB classifiers by removing attributes, one by one, from the least important (lowest ranked) attribute, until only the highest ranked attribute remains, e.g., each classifier is run with attributes A-F, then A-E, A-D and so on, which can be expressed as generally important attributes. Finally, the best subset of attributes, recording the highest classification accuracy in a single cross-validation fold on the training set, is selected as the final candidate attribute set, to be tested on the test set (10%) via each generated classifier to provide the final classification accuracy (step 3, in Fig. 2-1).

2.2.3. Attribute selection methods

This section describes the Tree Node Selection (TNS) method, followed by a brief introduction of the methods of IG, RLF, CFS, CNS and WRP. Full descriptions of these methods are in WEKA (2005). TNS is a ranking filter AS method, as previously discussed. TNS repeatedly evaluates and removes the lowest ranked attribute until no attributes remain.

The best performing attribute set will be chosen and will be tested by pruned and unpruned C4.5 decision tree and NB classifiers (Fig. 2-1). However, the selection process of TNS is different from any other AS method. TNS selects the most predictive attributes by considering each of the decision nodes in the pre-generated pruned and unpruned decision trees from a training set. This can be considered an alternative to *pruning,* as previously discussed. TNS counts the number of instances that are classified by paths passing through each node or leaf node to assess the significance of each attribute. As like other AS methods, the purpose of TNS is to select fewer attributes, retaining the most predictive, to improve results (classification accuracy) without losing meaning or altering the nature of the problem. A unique approach in TNS is to explore both pruned and unpruned decision trees to select predictive attributes, called TNSP and TNSU, respectively. A motivation of using both pruned and unpruned decision trees is that practically not all pruned trees obtain better classification accuracy than the corresponding unpruned tree, even though trees are pruned in order to obtain better classification accuracy. Unpruned decision trees can be over-fitted and difficult to interpret due to their large tree size, thus their practical use is not common. However, unpruned decision trees are still informative, as in fact, they select many predictive attributes that are worth investigating (Freitas 2002). The TNS method is described as follows:

2.2.3.1. Tree Node Selection algorithm

Let $T = (V, F, E, Lv, Lf)$ be a generated decision tree (Fig. 2-2). The nodes are represented as $V(T) = \{v_1, \ldots, v_{nv}\}$, where nv is the total number of nodes in the decision tree T (excluding leaf nodes). Let A be a set of input attributes where $A = \{a_1, \ldots, a_{na}\}$ and na is the number of attributes. The labels corresponding to the nodes in $V(T)$ are represented as $Lv = \{L(v_1), \ldots, L(v_{nv})\}$ and $L(v_i) \in A \; \forall \; v_i \in V(T)$, where $L(v_i)$ is the label for node v_i.

Not all attributes need be used. For example (Fig. 2-2 right), when there are four input attributes ($na = 4$), only two attributes and three nodes ($nv = 3$) might be used to construct T, so $V(T) = \{v_1, v_2, v_3\}$ and the corresponding labels are $Lv(T)$ could be $\{a_1, a_2, a_2\}$, indicating that node v_1 is labelled with attribute a_1, and nodes v_2 and v_3 are labelled with the same attribute, a_2.

Similarly, the leaf nodes are represented as $F(T) = \{f_1, \ldots, f_{nf}\}$, where nf is the number of leaf nodes. Hence, the size of the decision tree (T) is $nv + nf$. Let C be a set of classes where $C = \{c_1, \ldots, c_{nc}\}$ and nc is the number of classes. The labels corresponding to the leaf nodes $F(T)$ are represented as $Lf = \{L(f_1), \ldots, L(f_{nf})\}$, and $L(f_i) \in C \; \forall \; f_i \in F$, where $L(f_i)$ is the label for leaf node f_i. For example, if there are two input classes ($nc = 2$; class c_1 for *yes* and c_2 for

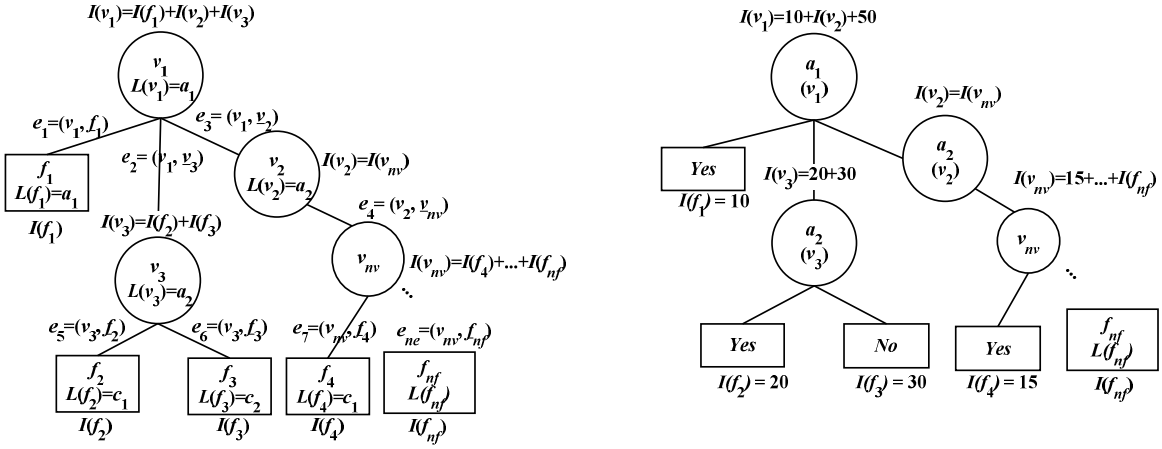


Fig. 2-2 Description of Tree Node Selection process (left) and an example of the decision tree (right).

no) and four leaf nodes were created as $F(T) = \{f_1, f_2, f_3, f_4\}$, the corresponding labels for $F(T)$ might be $L(f_i) = \{c_1, c_1, c_2, c_1\}$, which indicates that leaves f_1, f_2 and f_4 are labelled with the class *yes* (c_1), and f_3 is labelled with the class *no* (c_2) shown in Fig. 2-2 (right) as an example.

Connections between pairs of nodes (including leaf nodes) are represented by edges, $E(T) = \{e_1, \ldots, e_{ne}\}$ where ne is the number of edges in T. An edge e_i between two nodes (v_j and v_k) is defined as $e_i = (v_j, v_k) \mid v_j, v_k \in V(T)$, and an edge e_i between a node v_j and leaf node f_k is defined as $e_i = (v_j, f_k) \mid v_j \in V(T), f_k \in F(T)$.

Let I be the total number of correctly classified instances at a node or leaf node, such that $I(f_i)$ represents the number of correctly classified instanced at leaf node f_i, and $I(v_i)$ is defined recursively,

$$I(v_i) = \begin{array}{ll} \Sigma\, I(f_j) & \forall\ f_j \mid (v_i, f_j) \in E \\ + \Sigma\, I(v_k) & \forall\ v_k \mid (v_i, v_k) \in E. \end{array} \qquad (2\text{-}1)$$

Note that I is calculated at a leaf node from the number of classified instances minus the number of incorrectly classified instances, in the output from WEKA.

For example,

$$I(v_1) = \begin{array}{ll} I(f_1) & \text{where } j=1,\ (v_1, f_1) \in E \\ + I(v_2) & \text{where } k=2,\ (v_1, v_2) \in E \\ + I(v_3) & \text{where } k=3,\ (v_1, v_3) \in E. \end{array} \qquad (2\text{-}2)$$

The number of instances classified by paths including node v_3 is calculated as $I(v_3) = 20 + 30$, as $I(f_2) = 20$ and $I(f_3) = 30$. Then, $I(f_1) = 10$, $I(v_2) = I(v_{nv})$ and $I(v_3) = 50$, so the total number of instances classified by paths including node v_1 is calculated as $I(v_1) = 10 + I(v_{nv}) + 50$.

Finally, each attribute a_i is ranked with the overall total instances, calculated from the total number of instances at nodes labelled with a_i,

$$I(a_i) = \quad \Sigma\, I\,(v_k)\ \forall\ v_k \mid L(v_k) = a_i. \tag{2-3}$$

A higher ranked attribute a_i has a larger $I(a_i)$, indicating that the attribute is more frequently used in the decision tree T, such that more instances are classified by rules involving a_i to determine the class. Note that the attribute ranking is based on correctly classified instances, but attributes can also be ranked using total or incorrectly classified instances.

The TNS algorithm is implemented as a Python program (using Python 2.5), which uses WEKA, written in Java.

2.2.3.2. Information Gain

Information Gain (IG) is one of the simplest and fastest attribute selection methods (Hall and Holmes 2003). IG requires discretised data and is a ranking filter method. IG selects and orders attributes by importance by measuring the information gain with respect to the class;

$$\textit{InformationGain}\ (\textit{Class}, \textit{Attribute}) = H(\textit{Class}) - H(\textit{Class} \mid \textit{Attribute}),$$

where $H(\textit{Class} \mid \textit{Attribute})$ is the conditional or best hypothesis of *Class* under *Attribute* from some space H, details in Quinlan (1993), Mitchell (1997) and WEKA (2008).

2.2.3.3. Relief

Relief is a ranking filter method, developed by Kira and Rendell (1992) to be robust, tolerate incomplete and noisy data, and manage multiclass problems. Its extension ReliefF was later developed by Kononenko (1994). Relief does not require discretised data, and is applied for both discrete and continuous classes as a pre-processing step and can be used to select splits in the building phase of decision tree during the learning process (Kononenko et al. 1997). Relief searches for important attributes by repeatedly selecting a randomly selected instance from its two nearest neighbours between the same class and a different class, and updating the quality estimation for all the attributes (details in Kira and Rendell 1992; Kononenko 1994; Kononenko et al. 1997). The main parameters for Relief are m (the number of instances sampled) and k (number of nearest neighbours). Kononenko (1994) recommended that a larger the m value provides better and reliable estimation. In this study, to obtain the best performance fairly among all other methods, the m is set to sample all instances and k=10 (default setting for WEKA).

2.2.3.4. Wrapper subset evaluator

Wrapper, WRP, does not require discretised data and is a subset attribute selection method. WRP uses a user selected learning classifier to evaluate attribute sets and estimates the accuracy of the learning for a set of attributes via cross validation (Hall and Holmes 2003). In this study, the identical test classifier, pruned C4.5 decision tree (C4.5-P), unpruned C4.5 decision tree (C4.5-U) and NB, respectively, is selected for WRP, and five-fold cross validation is selected for accuracy estimation (default setting of WEKA).

2.2.3.5. Correlation-based Feature Selection

Correlation-based Feature Selection, CFS, requires discretised data, and is a subset attribute evaluator, developed by Hall (1998); see details of the algorithm in Hall (1998, 2000). CFS is a subset evaluation heuristic, where a good subset of attributes is detected by considering the usefulness of individual features at predicting each class along with the level of intercorrelation (dependency) among them. Preferably, subsets of features are highly correlated with classes but their intercorrelation is low (Hall and Holmes 2003; WEKA 2008).

2.2.3.6. Consistency-Based Subset Evaluation

Consistency-Based Subset Evaluation, CNS, requires discretised data, and is a subset attribute evaluator. WEKA uses CNS based on Liu and Setiono (1996), which searches for a subset of attributes with the best consistency in the class values (Hall and Holmes 2003; WEKA 2008). Note that the greedy stepwise method (forward selection search, default setting of WEKA) is used for all the subset evaluators, CFS, CNS, and WRP.

Note that all of attribute selection methods described in this chapter are heuristic methods, i.e. they are not guaranteed to give the best solution, but will generally give one that is close to optimal (Baase and Van Gelder 2000).

2.2.4. Assessment on selected attributes using statistical analysis

In this study, Analysis of Variance (ANOVA), Tukey's test and the Kruskal-Wallis test for multiple comparisons were applied to assess three performances metrics: difference of classification accuracy before (without AS) and after AS (with AS), relative reduction of attributes and relative reduction of tree size, for each method over three classifiers: pruned decision tree (C4.5-P), unpruned decision tree (C4.5-U) and naïve Bayes (NB). All statistical analyses were carried out using Minitab 15.1 (Minitab 2008). The difference of classification accuracy (CA difference) is calculated from the classification accuracy after AS minus before AS, to show its improvement, i.e., positive CA difference values indicate an improvement

with AS. Relative reduction (RR) of attributes is calculated from the final number of attributes selected by AS divided by the original input number of attributes and multiplied by 100%, then subtracted from 100%, i.e., a large value (in %) indicates a large reduction of attributes. Relative reduction (RR) of tree size is calculated from the final tree size divided by the original tree size and multiplied by 100%, then subtracted from 100%, i.e., a large value (in %) indicates that a smaller decision tree is produced after AS. Note that this observation is only applicable for C4.5.

The overall mean and standard deviation ($\mu \pm SD$) values of CA difference, RR of attributes and RR of tree size are calculated from the 10-fold cross validation (n=10) separately for before (without AS) and after AS (with AS) for three classifiers; C4.5-P, -U and NB. In order to assess different attribute selection methods effectively and fairly, firstly, analysis of variance (ANOVA) with 10-fold replication with 95% confidence intervals (CI) is used to test whether there is significant evidence of an effect of methods and data. Tukey's test was applied to identify which methods were significantly different from the others.

Tukey's test examines all pairwise differences of means among AS methods by controlling the family error rate (set as 0.05 for 95% CI) but using smaller values of α for each individual confidence interval than the generally used CI, in this case 95%, to ensure that the confidence interval contains the true difference of all means. Tukey's method was used because undertaking a sequence of multiple comparisons can inflate the overall Type I error (McClave and Sincich 2003). Tukey's test is interpreted by Bon grouping, which connects methods that are not significantly different in their means. The highest ranking is labelled from 'A', 'B' and so on. If AS methods share the same Bon grouping letter, this indicates that the mean distance of these methods are not significantly different, i.e., the mean performance is similar among these methods. Sharing more than one Bon grouping letter indicates the mean distance overlaps with another method. For example, a Bon grouping 'A' was assigned for Method I, 'A' and 'B' were assigned for Method II, and 'B' was assigned for Method III. This suggests that Method I and Method III are respectively ranked the highest and lowest, and their mean distances are significantly different as they do not share a Bon grouping letter. However, Method I and Method II are not significantly different in their means, because the mean distance of Method II overlaps with both other methods. It suggests that the rank of Method II lies in the middle, and its performance quality was not significantly different from either method. Tukey's test was used to investigates the mean performance of CA difference, RR of attributes and tree size among AS methods, but a pair of each data set

for Tukey's test became 528 combinations for data, which makes interpretation impractical. Thus, the following section mainly discusses performance among methods.

A non-parametric method, the Kruskal-Wallis (KW) test, is applied to test the equality of medians for methods. The KW test also ranks performance of CA difference, RR of attribute and tree size among methods over all classifiers. It tests the hypotheses of whether the population medians are all equal or not. When a median value is considered, the independent absolute value is directly compared among methods. This is like looking at win-loss relationships among methods. Rankings of the KW test can be fair when the performance of the method shows high standard deviation values for the particular data, e.g., one fold provided the highest value and another fold provided the lowest value, since observation of the mean value cancels out these effects. The KW test provides median, mean ranking, and z-value, which represents the significance of its ranking. A large positive z-value generally has a large mean ranking value, which indicates the mean ranking is different from all observations, and shows positive improvements after AS. A small absolute z-value indicates least difference from all observations, and a negative z-value, smaller mean ranking value, shows least improvement after AS process. Note that the large values of CA difference, RR of attributes and tree size indicate the improvement with AS. Here, the z-value is ranked from large to small z-value, as 1 to 7, such that rank 1 indicates the largest improvement. Generally, the ranking of the KW test agrees with the ranking of means, but if not, the KW test is used to report the final ranking in this study.

In addition to the above statistical tests, it is interesting to examine the performance differences between pruned and unpruned decision trees. The paired *t*-test is carried out to test the overall mean difference of CA difference, RR of attributes and RR of tree size over 33 datasets (n=33) between paired observations for pruned and unpruned decision tree among the same attribute method using 95% confident intervals. Lastly, all ANOVA tests were carried out with replication. This means, for example, that the ANOVA test was investigated on a difference of classification accuracy before and after AS within one fold over 33 data for 10-fold (n=330).

2.2.5. Processing time

Two different processing times for a single fold of cross-validation will be reported, for an Athlon XP 2800+ with 1GB RAM. Firstly, the processing time (in seconds) for the entire process is reported for all methods. For TNS, this includes the time to generate the C4.5 decision trees (Step 1 in Fig. 2-1) that TNS takes as input, selecting attributes via TNS, and evaluate the subsets of attributes (Step 2 in Fig. 2-1), whereas for subset evaluator

approaches, such as CFS, the time taken to select the attribute subset (Steps 1-2 in Fig. 2-1) is taken. Secondly, to compare the ranking filter approaches, TNS, IG and RLF, the time to generate the ranked attributes (Step 1 in Fig. 2-1), excluding the evaluation step, is assessed. The processing times are then investigated by One-way ANOVA for all methods, subset methods, and ranking filter methods.

2.3. Results and discussion

This study investigated commonly known classifiers, C4.5 and naïve Bayes, to test performance of the newly developed Tree Node Selection (TNS) method and compared with five known AS methods: IG, RLF, CFS, CNS and WRP, using 33 benchmark datasets with various ranges of instances, attributes and classes. TNS investigates the pre-generated decision tree by assessing nodes in both pruned and unpruned decision trees by counting frequencies of instances that are classified by paths involving certain nodes as an alternative to *pruning*, and rank them for individual attributes by their importance. TNS has two approaches, TNSP and TNSU, which use respectively pruned and unpruned decision trees as their information source. Ideally, the best attribute selection method is to reduce the highest proportion of irrelevant attributes effectively from data as well as improve results in classification accuracy, number of attributes and tree size without losing the nature of the data or problem.

The overall mean and standard deviation ($\mu \pm SD$) values (10-fold cross validation, n=10) of classification accuracy, number of attributes and decision tree size for each data set performed by all classifiers without AS and with seven AS methods are shown respectively for each classifier, C4.5-P, -U and NB, in Appendix 2-1 to 2-8. Table 2-2 (top) shows a summary of overall mean and SD values over 33 datasets of classification accuracy, number of attributes and tree size, and separately calculated means and SD values of difference of classification accuracy (CA difference), relative reduction (RR) of attributes (%), and relative reduction (RR) of decision tree size (%) before and after attribute selection (AS) for each method are shown in Table 2-2 (bottom). Direct outputs of each statistical analysis are shown in Appendices; ANOVA (Appendix 2-9), Tukey's test (Appendix 2-10) and Kruskal-Wallis test (Appendix 2-11) for CA difference, RR of attributes and tree size before and after AS. Note that ANOVA tests were carried out to test for equality or if at least two means are different at 95% CI, and Tukey's test individual CI was 99.68% CI for each method and 99.98% CI for data to control experiment-wise error rate for multiple comparisons of a set of means.

Table 2-2 Overall mean and standard deviation values of classification accuracy over 33 data sets.

Assessments	Test classifier	Original*	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
Classification accuracy (%)	C4.5-P	82.45 ± 13.09	82.65 ± 12.74	82.29 ± 13.13	81.99 ± 12.58	82.06 ± 12.35	82.98 ± 12.44	82.92 ± 12.41	80.02 ± 13.74
	C4.5-U	81.84 ± 13.40	82.52 ± 12.70	82.22 ± 13.12	81.64 ± 12.92	81.50 ± 12.58	82.97 ± 12.61	82.65 ± 12.51	79.45 ± 13.85
	NB	79.37 ± 14.60	80.80 ± 13.24	80.92 ± 13.36	78.95 ± 14.40	79.67 ± 14.17	80.78 ± 14.32	81.20 ± 14.02	77.11 ± 14.58
Original attribute number	C4.5-P	32 ± 47	7.29 ± 5.07	7.90 ± 5.64	5.57 ± 3.24	8.52 ± 5.74	11.95 ± 11.14	11.50 ± 7.61	9.02 ± 4.60
	C4.5-U	32 ± 47	7.21 ± 5.36	7.56 ± 6.38	5.57 ± 3.24	8.52 ± 5.74	11.68 ± 13.09	10.75 ± 8.96	9.02 ± 4.60
	NB	32 ± 47	7.46 ± 5.93	8.63 ± 6.74	5.57 ± 3.24	8.52 ± 5.74	10.67 ± 8.49	11.14 ± 8.22	9.02 ± 4.60
Original decision tree size	C4.5-P	87.9 ± 222.3	75.2 ± 209.5	73.6 ± 207.7	65.7 ± 202.4	66.0 ± 202.6	78.3 ± 213.4	81.8 ± 213.3	72.9 ± 208.1
	C4.5-U	453.2 ± 1540.9	84.9 ± 229.4	101.8 ± 245.4	124.2 ± 343.6	102.1 ± 236.6	422.3 ± 1434.7	335.6 ± 1420.4	393.1 ± 1233.7
Difference of classification accuracy before and after AS (%)	C4.5-P	-	0.20 ± 3.95	-0.16 ± 4.03	-0.46 ± 6.58	-0.39 ± 4.24	0.53 ± 4.11	0.47 ± 4.45	-2.43 ± 9.26
	C4.5-U	-	0.68 ± 5.12	0.39 ± 4.75	-0.19 ± 6.86	-0.34 ± 5.20	1.13 ± 5.13	0.81 ± 5.06	-2.39 ± 9.04
	NB	-	1.43 ± 7.59	1.54 ± 7.40	-0.42 ± 7.85	0.30 ± 6.56	1.41 ± 5.37	1.83 ± 7.12	-2.27 ± 10.72
Relative reduction of attributes (%)	C4.5-P	-	66.4 ± 22.5	63.6 ± 24.3	72.8 ± 19.0	61.2 ± 20.4	52.8 ± 26.7	51.1 ± 27.0	54.0 ± 26.6
	C4.5-U	-	68.1 ± 20.3	67.4 ± 22.5	72.8 ± 19.0	61.2 ± 20.4	56.8 ± 26.8	56.0 ± 26.8	54.0 ± 26.6
	NB	-	65.2 ± 24.5	60.5 ± 25.2	72.8 ± 19.0	61.2 ± 20.4	53.8 ± 28.8	50.4 ± 29.1	54.0 ± 26.6
Relative reduction of decision tree size (%)	C4.5-P	-	14.3 ± 48.4	19.0 ± 31.2	-8.0 ± 242.0	15.3 ± 42.8	12.7 ± 31.2	-116.9 ± 816.1	6.7 ± 58.4
	C4.5-U	-	35.9 ± 35.9	33.2 ± 36.4	42.5 ± 39.1	25.8 ± 34.0	10.4 ± 128.4	24.8 ± 34.2	0.7 ± 61.2

* Using the test method without attribute selections.

In comparison to this study, Hall and Holmes (2003) used 15 benchmark datasets (plus three large datasets) to investigate the same AS methods as this study; CFS, IG, CNS, RLF and WRP, to test the classification accuracy, number of attributes and tree size using C4.5 pruned and NB classifiers. However, the statistical approaches that were taken by Hall and Holmes (2003) make this study incomparable. For example, they only reported the mean values of 10-fold cross-validation (n=10), and all AS methods were assessed by a paired t-test (two-tail) with one percent significance level, counting how often each method performs significantly better or not. Separately, the performance of AS methods were ranked by the total number of "wins" minus "losses" that were counted from the number of times each method is significantly more or less accurate than another before and after AS is performed.

During this study, it was found that combinations of different but less than 20 benchmark datasets provided unstable p-values among methods, such that the best performing AS methods took different places and p-values indicated the means among methods were not significantly different for some cases (at α=0.05). However, the analysis became stable to provide strong evidence (p-value < 0.05) after at least 30 benchmark datasets were tested. Also, it is not critical, but is important to consider the mean and standard values (and 95% CI) together to assess the performance difference among data and among methods, as the following section describes.

2.3.1. Overall mean and standard deviations of attribute selection experiments

Observed from overall mean and SD values for each method in Table 2-2, each fold was fairly generated for the test, but standard deviation values vary within each dataset. This indicates that some folds do better or worse within data, for example, *labor-relations* data shows the high standard value of classification accuracy (85.0±17.48 in Appendix 2-1) for C4.5-P without AS. This suggests that the overall estimation of the performance from such particular data, e.g., *labor-relations*, is less precise than others that have lower SD values and narrower CI.

2.3.2. Two way-ANOVA tests

Outputs of Two-way ANOVA are shown in Appendix 2-9. All three assessments (CA difference, RR of attributes, and RR of tree size) for all three classifiers (C4.5-P, U and NB) have p-value of < 0.001 for factor interaction, main effect of AS methods and data, respectively. These indicate that at least two means of AS methods or data are significantly different, and AS methods and data interact to affect the performance of CA difference, RR of attributes or tree size measurements. These results satisfy to test a multiple comparisons

procedure, One-way ANOVA and Tukey's test, and Kruskal-Wallis test, to compare all pairs of method or data means in CA difference, RR of attributes and tree size.

2.3.3. Tukey's and Kruskal-Wallis test

All results from One-way ANOVA show p-value of < 0.001 separately among methods and data (note that outputs of One-way ANOVA results are not shown, as SS and MS values can be referred from Two-way ANOVA in Appendix 2-9). This suggests rejecting the null hypothesis that all means of AS methods or data are equal that there is enough evidence to say that at least two means of AS methods or data are different. Hence, Tukey's test was carried out. Output values of Tukey's test, all pair wise difference of CI for means that can be used for Bon grouping are shown in Appendix 2-10, and detailed investigations were carried out for AS methods. Similarly, output values of the KW test for AS methods are shown in Appendix 2-11. All results of the KW test also have p-value of < 0.001 for both hypotheses, with or without adjusting for ties. Results suggest rejecting the null hypothesis that the population medians that are all equal. Thus, there is evidence to say that CA difference, RR of attributes and RR of decision tree size can be ranked among AS methods by their median values.

2.3.4. Interval plots for interpreting results

In order to make interpretation easier, interval plots are drawn to show combined results of means, 95% CI for means, Tukey's and the KW-test, separately for each performance in CA difference (Fig. 2-3), RR of attributes (Fig. 2-4) and RR of tree size (Fig. 2-5) for each AS method respectively for three classifiers, C4.5-P (left), C4.5-U and NB (right) in each plot. Lines in interval plots indicate the position of the lower and upper 95% CI for the mean, the number at the top of each line indicates the mean value, the letter under each line indicates the Bon grouping letter for Tukey's test, a number under each letter indicates the rank for the KW test based on the z-value (the number inside the brackets shows the z-value, and underlined z-values indicate reasonably large z-values).

The following sections describe assessments of each performance in CA difference, RR of attributes and RR of decision tree size among AS methods to examine how and which AS method performed significantly importantly over different classifiers.

2.3.5. Assessment of classification accuracy improvement with AS

The overall means of classification accuracy difference before and after AS show different improvements among methods over different test classifiers. From observing the highest positive means, sharing the best Bon grouping 'A', and relatively higher positive z-value of

the KW test ranking (z-value above 2.0, underlined values in Fig. 2-3), the best three performances are found from IG, followed by RLF and TNSP for both C4.5-P and –U, and the best four performances for NB are found from TNSU, followed by RLF, IG and TNSP.

The mean classification accuracy performances of all AS methods (except CNS) did not significantly differ for C4.5-P, since all AS methods except CNS have a single Bon grouping 'A'. For example, the top three mean improvements before and after AS methods were 0.53% for IG, 0.47% for RLF, 0.20% for TNSP, but the mean performance of CNS was consistently found significantly lowest for all classifiers, e.g., μ=-2.43 for C4.5-P and a single lowest Bon grouping for all classifiers. This suggests that TNSP performed similarly to existing AS methods. All AS methods (except CNS) select predictive attributes that can improve on or approximately retain the quality of a pruned decision tree rule rather than dramatically improve the prediction ability, since the maximum overall mean improvement difference in classification accuracy is 0.53%, observed from IG. This also suggests that the C4.5 and NB algorithms themselves select good attributes to build rules, so that the selection of attributes by the classifier itself and the AS method can be similar (except CNS). From here, the cost effective data collection and management is possible via AS methods, as AS methods identify fewer and good attributes for the model without degrading the quality of the model, in particular for the most reliable classifier, C4.5-P. However, the slightly less reliable classifier,

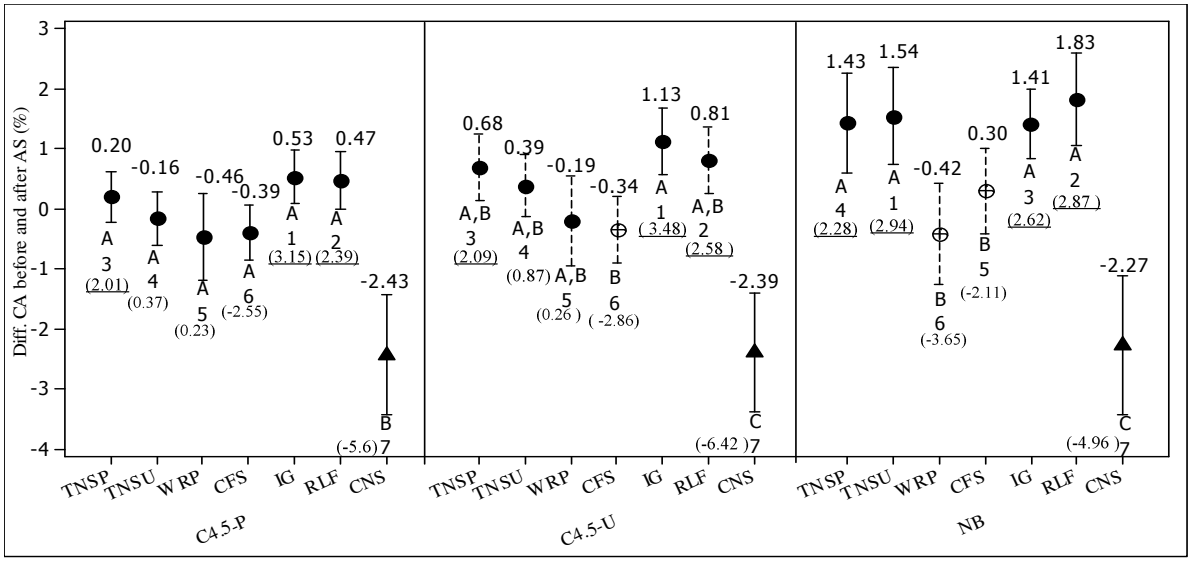


Fig. 2-3 Interval plots for differences of classification accuracy (CA) before and after attribute selection (95% CI for the mean).

The three sections show, from left to right, results of C4.5 pruned, C4.5 unpruned and NB classifiers. Each section shows the performance of seven AS methods from TNSP to CNS. Each plot shows the mean value of CA (top of 95% CI plot and black dot), the CI intervals (lines), Bon grouping letter (Tukey's test), the rank of KW test, and z-value of KW (in brackets). The top three z-values of the KW test are underlined.

C4.5 without pruning, and a different algorithm, NB, show slightly different improvements among AS methods.

For C4.5-U, the performance of IG (μ=1.13) stands out, since IG has a single Bon grouping 'A', followed by RLF (μ=0.81), TNSP (μ=0.68), TNSU (μ=0.39), WRP (μ=-0.19) and CFS (μ=-0.34). The second and third best AS methods, RLF and TNSP, overlapped a Bon grouping 'B' with the lower ranked methods, WRP and CFS. Note that the KW test ranking agreed with the order of means.

For NB, the best Bon groupings and higher KW test ranking, the best performance is ranked as TNSU, followed by RLF, IG and TNSP whereas the highest means are detected from RLF (μ=1.83), TNSU (μ=1.54), IG (μ=1.41) and TNSP (μ=1.43). The KW test ranking and the order of means slightly disagreed for the best and second performance, between TNSU and RLF that TNSU produced the highest median values and RLF produced the highest means.

It is reasonable to expect that TNS perform well on C4.5, as the information source of TNS is C4.5, and it is the same for WRP. Generally, WRP is known to provide good results due to its inductive algorithm. However, the overall highest mean of classification accuracy difference for NB was observed from both ranking filter methods; TNSU and RLF (which achieved similar best results). It could be suggested that TNS searches for attributes by ranking them from the highest frequency of instances that are classified by paths including certain nodes (including leaf nodes), TNS inputs might have a higher conditional probability, $P(A \cap B)$, among pairs of attributes. RLF selects instances from the two nearest neighbors between the same class and a different class. It could be possible that NB, based on Bayes theorem, may work better with such conditions to bring the high probability among selected attributes (strengthened probability), but further investigation will be required to conclude this.

As NB assesses all attributes every time, irrelevant attributes in the input data add noise to the classification, which can result in lower classification accuracy. Therefore, when TNS is found to perform best in NB, it may suggest that TNS must have selected the most relevant attribute set for NB.

Another point from the TNS approach is that the good use of unpruned decision trees is possible, since overall better mean classification accuracies are observed from TNSP for both C4.5 pruned and unpruned classifiers. It suggests that C4.5 with pruning is effective itself in pre-selecting good predictive attributes. Further, TNS ranked the attributes and used only the fewest good predictive attributes for further C4.5 pruned and unpruned decision tree

Table 2-3 Summary results of p-value (two-tail) of t-test for paired two sample for means* of pruned and unpruned.

AS method	CA difference	R.R. attribute	R.R. tree size
C4.5	**0.030**	-	0.154
TNSP	0.634	0.799	0.337
TNSU	0.805	0.529	0.101
WRP	0.052	-	0.162
CFS	**0.018**	-	**0.010**
IG	0.971	0.596	0.162
RLF	0.284	0.185	0.281
CNS	0.136	-	0.127

Diff = Pruned-Unpruned (n =33)

*Input values are a mean value of each data gained from 10-fold cross validation (n=10).

construction. In fact, even though the pruning process is generally known to result in better classification, it can also decrease classification accuracy by over-pruning. The pruning process uses a test set to test for attributes that do not appear to add to the classification accuracy, which will often contain examples that are not seen in the training set; if these examples result in errors, associated attributes may be pruned incorrectly.

Interestingly, all classifiers selected the same best three AS methods; IG, followed by RLF, TNS (TNSP for C4.5 and TNSU for C4.5), in Fig. 2-4. The magnitude of the mean classification accuracy improvements with AS methods without pruning is much larger than with pruning. TNSP is more than three times better for C4.5-U (μ=0.68) compared with C4.5-P (μ=0.20), IG is almost two times better for C4.5-U (μ=1.13) than C4.5-P (μ=0.53), and RLF for C4.5-U (μ=0.81) is better than with C4.5-P (μ=0.47). It indicates that AS methods are more effective on unpruned trees (compared with unpruned trees generated without AS) than on pruned tree (compared with pruned trees generated without AS), since the mean CA improvement is much larger for unpruned trees, which is a reasonable observation. However, does this suggest that the AS process using TNSP, IG and RLF makes a significant difference in terms of improving classification accuracy between generating pruned with AS methods and unpruned with AS methods? The *t*-test for paired two samples for means of both pruned and unpruned classification accuracy after AS in Table 2-3 shows that the means of classification accuracy of CFS (p=0.018) and C4.5 without attribute selection (p=0.030) are the only significant differences between pruned and unpruned decision trees (the higher mean was detected from pruned, and actual mean difference is shown in (Appendix 2-13). This suggests that the default pruning approach effectively provides better classification accuracy for C4.5 than not pruning (without any AS process). However, when AS processes TNS, WRP, IG, RLF, or CNS (but not CFS) were applied to construct a decision tree, the mean classification accuracy improvement is insignificant whether pruned or unpruned decision tree classifiers were used. The mean classification accuracy using CFS provided improved pruned prediction than unpruned, but its overall mean classification accuracy improvement is generally lower than IG, RLF and TNS.

The ranking filter approach is said not to be as effective and general as subset evaluation, as it ignores the effect of subsets of attributes in the performance of the induction algorithm (Schuschel and Hsu 1998), but the best three common AS methods for all classifiers that improve or sustain the classification accuracy are all ranking filter methods, IG, RLF and TNS, where the preferable AS method is one that improves or sustains classification accuracy as well as selects fewer attributes. Thus, the following section will examine whether IG, RLF and TNS are satisfactory or not.

2.3.6. Reduction of attributes among AS methods

The performance on attribute reduction is assessed similarly to the previous section by referring to the combined statistical information in interval plots, shown in Fig. 2-4. While IG, RLF and TNS have the best performance to improve classification accuracy, only TNS is found to select fewer attributes among them. Firstly, WRP is found to have the best performance of selecting significantly fewer mean proportion of attributes over all classifiers (overall means, μ, of 72.8% of attributes are removed for all data), followed by TNSP (μ=66.4 for C4.5-P, μ=68.1 for C4.5-U and μ=65.2 for NB), and TNSU (μ=63.6 for C4.5-P, μ=67.4 for C4.5-U and μ=60.5 for NB). The performance of WRP over three classifiers

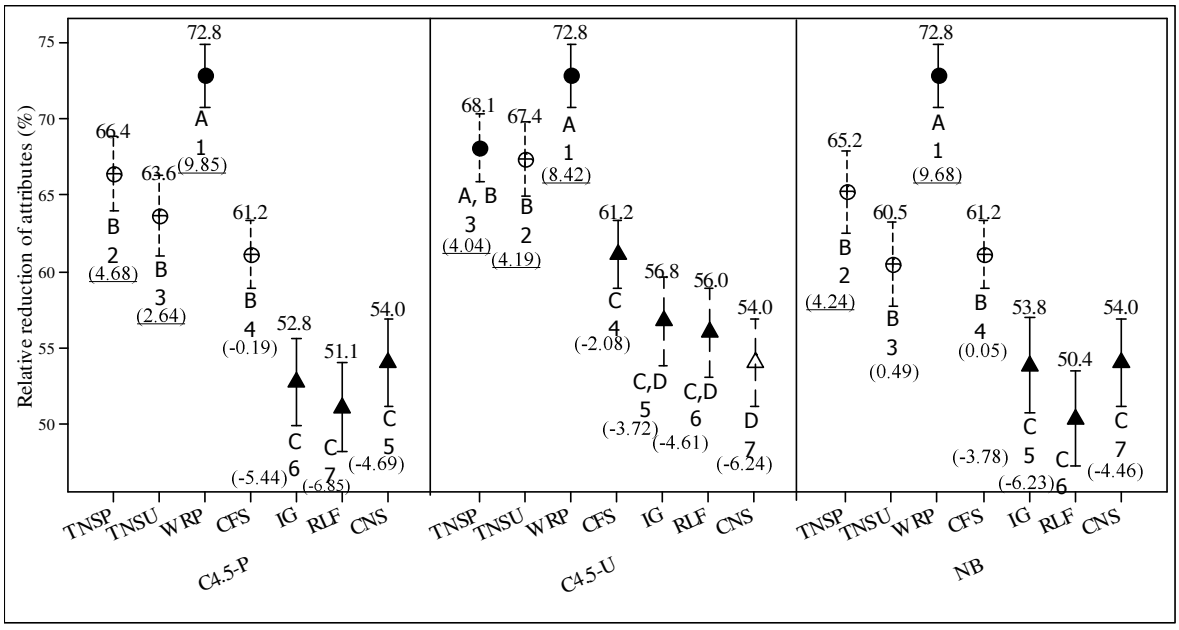


Fig. 2-4 Interval plots for relative reduction of attributes before and after attribute selection (95% CI for the mean).

The three sections show, from left to right, results of C4.5 pruned, C4.5 unpruned and NB classifiers. Each section shows the performance of seven AS methods from TNSP to CNS. Each plot shows the mean value of relative reduction (top of 95% CI plot and black dot), the CI intervals (lines), Bon grouping letter (Tukey's test), the rank of KW test, and z-value of KW (in brackets). The top three z-values of the KW test are underlined.

stands out compared with the second best performance of TNSP and the rest of the AS methods, since the mean distance and medium ranking between WRP and TNSP are significantly different for all classifiers, shown from the different Bon grouping letters assigned, 'A' for WRP and 'B' for TNSP, and extremely large KW test z-value was observed from WRP for all classifiers, e.g., the minimum z-value found is 8.42 for C4.5-U, compared with TNS, e.g., the maximum z-value is found as 4.68 for C4.5-P.

Generally, TNSP provided better results than TNSU by having the higher means of RR of attributes for all classifiers, e.g., the range of the mean values over three classifiers is 66.4-68.10% for TNSP and 60.5-67.4% for TNSU. TNSP has the higher KW test ranking with much higher positive z-values (z=4.68 for C4.5-P and z=4.24 for NB) than TNSU (z= 2.64 for C4.5-P and z=0.49 for NB), but TNSP and TNSU performed similar for C4.5-U, in fact the TNSU mean medium ranking was higher (z=4.19) than TNSP (z=4.04), though the mean distance of TNSP is closer to WRP than TNSU (TNSP for C4.5-U overlapped with WRP's Bon grouping 'A'). A possible scenario is that TNSP provides better results on reducing number of attributes than TNSU for all classifiers because the initial pruned decision tree already selected fewer attributes than unpruned, thus input attribute numbers were smaller for pruned than unpruned. However, note that TNSU provides better classification accuracy for NB than TNSP. Unpruned trees may select more attributes to construct the larger tree than pruned, thus the number of input attributes for TNS is larger than pruned; this can mean that chances of selecting better attributes will be increased for unpruned than pruned, that may help increasing selecting more highly intersected attributes for NB. However, further investigation is required for this.

Fourth best performance is observed from CFS (μ=61.2). The means of CFS performed similar to TNS since it is ranked in the same Bon grouping 'B' as TNS for C4.5-P and NB. However, CFS did not perform similarly to TNS for C4.5-U, since CFS belongs to another Bon grouping 'C', ranked the same as the lower mean performance, IG and RLF. IG, CNS and RLF could only remove about half of the attributes for all classifiers (the mean ranges or mean over three classifiers are 52.8-56.8 for IG, 54.0 for CNS, 50.4-56.0 for RLF). In fact, the KW test z-values of CFS, IG, RLF and CNS are small, e.g., the maximum z-value, 0.05, is observed from CFS for NB among these AS methods. This also suggests that these AS's median ranking values are significantly lower than WRP and TNS.

In spite of IG and RLF showing the best performance on classification accuracy, they were two of the lowest-ranked AS methods in terms of mean reduction of attributes, i.e., a relative proportion of about 50-60% of attributes is removed. IG and RLF successfully removed

irrelevant or redundant attributes to help the classifier algorithm to improve the classification accuracy, but their selection process may not be successfully selecting fewer but more predictive attributes, because their mean classification accuracy improvement is statically similar to TNS, when in fact TNS removed additionally 10% more of attributes than IG and RLF for all classifiers, i.e., the mean range over three classifiers is 65.2-68.1% for TNSP, 52.8-56.8% for IG and 51.1-56.0% for RLF. If the experimenter wants to run the experiment conservatively on reducing number of attributes, then IG and RLF are satisfactory, as the better classification is assured. However, if the loss of classification accuracy by choosing TNS is not critical, as the mean improvement among IG, RLF and TNS is not significantly different, then TNS would provide a tool for the cost effective data collection and management, as TNS selects fewer but most predictive attributes with improved classification accuracy for C4.5 and NB classifiers. Generally, WRP is known to provide better results due to the interaction between the search and the learning scheme's inductive bias (Hall and Holmes 2003), but WRP is a time consuming method that is not practical. WRP selected the fewest attributes, followed by TNS, but the mean classification accuracy of TNS was better than WRP for all classifiers.

Additionally, the *t*-test for paired two-sample for means of RR of attributes between pruned and unpruned with AS methods were found not to be significantly different for any AS method, i.e., the minimum p-value among all AS methods is observed as 0.185 for RLF in Table 2-3. This suggests that all AS methods remove similar proportions of attributes to construct pruned or unpruned decision trees. In other words, constructing either pruned or unpruned trees does not change the ability of selecting fewer attributes for all AS methods.

2.3.7. Reduction of decision tree size among AS methods

Where the fewest attributes were selected by WRP, followed by TNS and CFS, this seems to directly contribute to constructing smaller decision trees. Its reduction was more effective for unpruned than pruned, as expected. WRP reduced the best mean tree size of 42.5% for unpruned, but the mean reduction of tree size for pruned was negative (μ=-8.0), indicating degraded results, in Fig. 2-5. This suggests that WRP overall performed the best to construct the smallest pruned decision tree over 33 data, but its overall estimation is less precise, e.g., WRP must perform very well on some data, e.g., *sonar* data's C4.5-P tree size was 7.4±4.9 for WRP, whereas originally 14.5±1.7 and 13.1±3.8 was for TNSP (Appendix 2-7). At the same time WRP performed poorly on other data, e.g., *mushroom* data's C4.5-P tree size was 30.9±1.4, 24.0±0.0 for both original and TNSP (Appendix 2-7). WRP shows the wider 95% CI for unpruned (-34.08, 18.14) than pruned (38.26, 46.69), i.e., longer CI line is observed for

pruned, in Fig. 2-5 (left) and actual CI values are shown in Appendix 2-12. This suggests that WRP performance was more precise for unpruned than pruned.

The second best performance is observed from TNSP (μ=35.9 for unpruned, μ=19.0 for pruned), then TNSU (μ=33.2 for unpruned, μ=35.9 for pruned). In comparison to WRP, TNS performed more consistently over the 33 datasets for both pruned and unpruned in that narrower 95% CI, e.g., TNSU CI (15.68, 22.41) for pruned, was observed, visible as shorter CI lines in Fig. 2-5 and CI values in Appendix 2-12. From comparing the mean values, an interesting observation is that TNSP using a pruned decision tree as its information source performed better to construct a concise unpruned tree (μ=35.9 for C4.5-U) than TNSU (μ=33.2 for C4.5-U), but TNSU that used unpruned information was better to construct a concise pruned tree (μ=19.0 for C4.5-P) than TNSP (μ=14.3 for C4.5-P), although this difference is insignificant from considering their mean distances (they share the same Bon grouping 'A' in Fig. 2-5).

Lastly, the third and fourth best performance, but similar mean distances to TNS were found from CFS (μ=25.8 for unpruned, μ=15.3 for pruned) and RLF (μ=24.8 for unpruned, μ=-116.9 for pruned). Similar degraded results to WRP are observed for RLF, which had the extremely wide 95% CI for pruned (-204.93, -28.82) in Appendix 2-12.

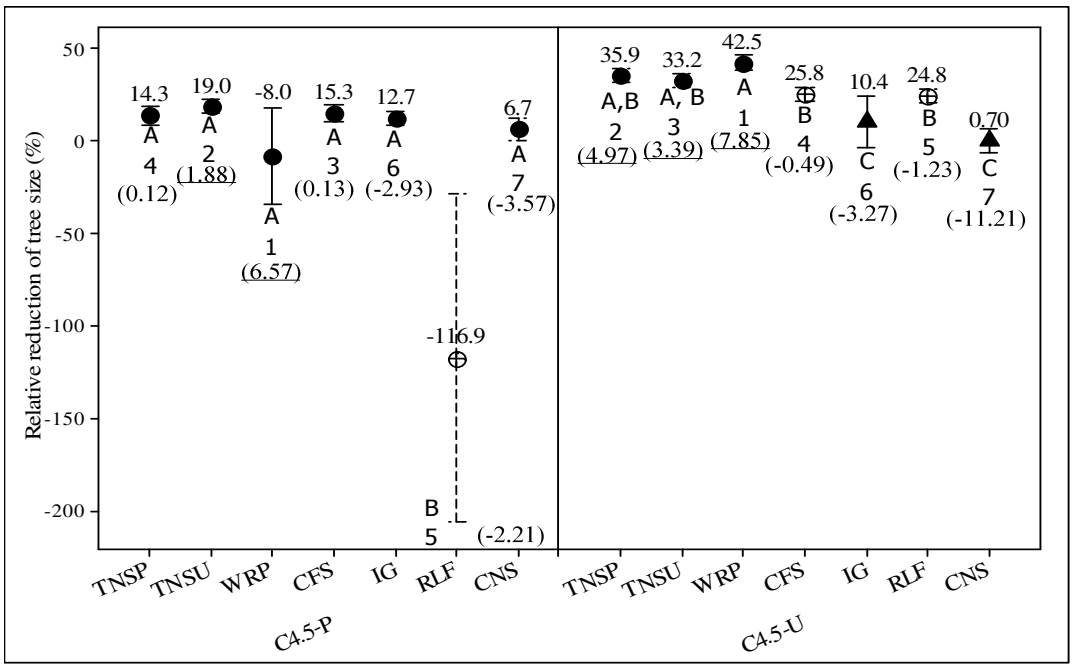


Fig. 2-5 Interval plots for relative reduction of decision tree size before and after attribute selection (95% CI for the mean).

The two sections show results of the C4.5 classifier, pruned (left) and unpruned (right). Each section shows the performance of seven AS methods from TNSP to CNS. Each plot shows the mean value of tree size reduction (top of 95% CI plot and black dot), the CI intervals (lines), Bon grouping letter (Tukey's test), the rank of KW test, and z-value of KW (in brackets). The top three z-values of the KW test are underlined.

Overall, WRP selects fewest and TNS (either TNSP or TNSU) selects the second fewest attributes among all AS, and WRP constructs the most and TNS constructs the second most concise tree for both unpruned and pruned. However, performance of TNS on classification accuracy is better than WRP for all classifiers in that TNS improved classification accuracy while selecting fewer attributes, but WRP more likely sustained classification accuracy for all classifiers, i.e., overall means of classification accuracy difference for WRP were all negative for all classifiers, but TNS has all positive means of classification accuracy difference for all classifiers except TNSU on C4.5-P. However, WRP and TNS performed similarly in classification accuracy for C4.5 (the same Bon grouping), but TNS performed significantly better than WRP for NB (different Bon grouping).

Additionally, the *t*-test for paired two sample for means of RR of tree size shows that only CFS shows that the mean reduction of tree size between pruned and unpruned is significantly different ($p = 0.010$ in Table 2-3). This suggests that CFS selects attributes that construct much smaller pruned than unpruned trees. Surprisingly, the overall mean difference of tree size between pruned and unpruned for original C4.5 was not significantly different ($p=0.154$ in Table 2-3). This may suggest that when a single dataset is observed, the pruned decision tree constructs a smaller tree than unpruned, but the overall mean difference between pruned and unpruned is not always significantly different when many different datasets are examined. Similarly, all AS methods except CFS select attributes that do not influence the mean tree size significantly whether a pruned or unpruned tree is constructed. As mentioned above, the performance of WRP and RLF vary among data for C4.5-P, since they have large 95% CI (Fig. 2-5).

2.3.8. Processing time for all AS methods

Results of the processing time (for a single fold) for all methods are shown in Table 2-4 for One-way ANOVA, and detailed statistical descriptions (in seconds) are shown in Table 2-5. Also, Appendix 2-14 shows the processing time for each data among all methods. Attribute selections labelled with P, U or N in brackets, indicates the test algorithm, C4.5 decision tree for pruned or unpruned or naïve Bayes, respectively.

The processing time for the ranking filter approaches including the attribute selection, starting from constructing the C4.5 decision tree for TNS, and evaluating each subset of attributes. Firstly, all mean processing times among the methods are not significantly different (p-value=0.117 in Table 2-4), thus the further statistical analyses are not carried out, e.g., Tukey's test. The best five individual attribute selection performance results (smallest mean value, in seconds) are found from CFS (0.6±0.8), followed by CNS (4.3±12.4), TNSP

Table 2-4 Outputs of One-way ANOVA test for all attribute selection methods for the processing time for a single fold (in seconds).

Source	DF	SS	MS	F	P
Factor	14	11050730	789338	1.47	0.117
Error	480	257475855	536408		
Total	494	268526586			

S = 732.4 R-Sq = 4.12% R-Sq(adj) = 1.32%

Table 2-5 Outputs of mean, standard deviation, individual 95% CI for pooled standard deviation for all attribute selection methods (*n*=33) for the processing time for a single fold (in seconds).

Each letter, P, U or N, inside the bracket indicates the method of testing algorithms, pruned and unpruned, or naïve Bayes, respectively.

AS Methods	Mean rank	Mean	StDev	Individual 95% CIs For Mean Based on Pooled SD
(P) TNSP	**8**	141.0	505.9	(--------*-------)
(U) TNSP	**6**	125.2	446.9	(--------*-------)
(N) TNSP	**3**	24.4	70.5	(--------*-------)
(P) TNSU	**9**	150.8	532.9	(--------*-------)
(U) TNSU	**7**	130.2	453.5	(--------*-------)
(N) TNSU	**4**	26.6	70.4	(--------*-------)
CFS	**1**	0.6	0.8	(--------*-------)
CNS	**2**	4.3	12.4	(--------*-------)
(P) IG	12	224.2	614.9	(--------*-------)
(U) IG	**10**	208.3	576.5	(--------*-------)
(N) IG	**5**	71.6	227.2	(--------*-------)
(P) RLF	14	367.6	1114.8	(--------*-------)
(U) RLF	13	355.0	1086.0	(--------*-------)
(N) RLF	11	214.8	711.6	(--------*-------)
WRP	15	548.0	1843.9	(--------*-------)

0 300 600 900

(24.4±70.5) and TNSU (26.6±70.4) when testing naïve Bayes, and IG when testing naïve Bayes (71.6±227.2). From the 6th to 9th performance is found from various types of TNS, followed by IG with pruned and unpruned C4.5 pre-generated decision trees, RLF, and the lastly WRP. As it is expected, the subset evaluator methods (excluding WRP), CNS and CFS, are efficient, compared with the ranking filter methods, TNS, IG and RLF. Among the ranking filters, TNS with naïve Bayes showed better performance than IG. Additionally, separate One-way ANOVA was tested for three subset evaluator approaches, but it is also found that the means among these three methods are not significantly different (p-value = 0.060, shown in Appendix 2-15. This suggests that even though CFS seems to perform extremely well, overall, the mean performance is not significantly different from other subset evaluator approaches and also among all other methods.

Interestingly, the processing times for the four filter approach methods, TNSP, TNSU, IG and RLF, excluding the evaluation time was similar (the processing time to generate the C4.5 decision tree to provide the initial attribute sets for TNS is included). Results of ANOVA, statistical description and the processing time for each data, are shown in Table 2-6, Table

Table 2-6 Outputs of One-way ANOVA test for attribute selection processing time for the ranking filter approach methods excluding evaluating the test algorithms, for a single fold (in seconds).

```
Source   DF      SS       MS      F      P
Factor    3    532894   177631   1.67  0.177
Error   128  13637166   106540
Total   131  14170060

S = 326.4   R-Sq = 3.76%   R-Sq(adj) = 1.51%
```

Table 2-7 Outputs of mean, standard deviation, individual 95% CI for pooled standard deviation for the ranking filter approach methods excluding evaluating the test algorithms, for a single fold (in seconds).

```
Level Mean   Mean  StDev    Individual 95% CIs For Mean Based on Pooled StDev
      Rank                -+---------+---------+---------+--------
TNSP   3     2.0    5.5    (----------*----------)
TNSU   2     1.9    5.0    (----------*----------)
IG     1     0.5    0.6    (----------*----------)
RLF    4   148.2  652.8                  (----------*----------)
                           -+---------+---------+---------+--------
                         -100         0       100       200
Pooled StDev = 326.4
```

2-7 and Appendix 2-16, respectively. The means of processing time among all filter approach methods are not significantly different (p-value=0.177 in Table 2-6), but the best performance was observed from IG (0.5±0.6), followed by TNSU (1.9±5.0), TNSP (2.0±5.5) and RLF (148.2±652.8), shown in Table 2-7.

Since TNS is written as a separate program in Python, while all other AS methods are a part of the WEKA software and written in Java, it is expected that some time is consumed waiting for the Java Virtual Machine to start and open the Java program each time an operation involving WEKA is executed, for example generating the initial decision tree, evaluating the candidate attribute sets, and testing. Even though this process is involved, One-way ANOVA test suggested the means among methods is not overall significantly different.

2.4. Conclusions

This study proposed a newly developed Tree Node Selection (TNS) method, which investigates the pre-generated decision tree as an information source to select the few most predictive attributes to help constructing further models. An interesting feature of TNS is to evaluate information from both pre-generated pruned and unpruned decision trees, called TNSP and TNSU, respectively. The performance of TNS was tested along with five known AS methods, CFS, CNS, IG, RLF and WRP, on 33 benchmark data sets with various numbers of instances, attributes and classes, using pruned and unpruned C4.5 decision tree and naïve Bayes classifiers. Results of classification accuracy, numbers of attributes and decision tree size before and after AS were measured, to examine the performance of each AS method.

TNS (TNSP and TNSU) is found to be the most consistent AS method over various data sets to select fewer attributes to construct smaller decision tree by improving the classification accuracy for C4.5 and NB classifiers. All other AS methods trade off the achievement between classification accuracy and the number of attributes. For example, the best achievements in classification accuracy are observed from IG and RLF, followed by TNS for C4.5 (pruned and unpruned), but IG and RLF removed the fewest attributes for all classifiers. WRP, followed by TNS (TNSP and TNSU) achieved the best, selecting fewest attributes, but WRP did not achieve better classification accuracy than IG, RLF and TNS for all classifiers. The uniqueness of TNS is to introduce the new use of unpruned decision trees. TNS could be expected to bias for C4.5, but the best performance of improving classification accuracy of NB was TNSU that uses information from an unpruned C4.5 tree (though RLF performed almost similarly to TNSU for NB). In comparison of TNSP and TNSU, TNSP generally shows better classification accuracy than TNSU for C4.5 (pruned and unpruned) and better reduction of attributes for all classifiers. Classification accuracy of C4.5 without any attribute selection is significantly better for pruned than unpruned, but the mean performance of classification accuracy, number of attributes and tree size was not significantly different whether a pruned or unpruned decision tree is constructed, if any AS methods except CFS were applied. Generally, various AS methods take top place for classification accuracy, reduction of attributes and reduction of tree size for different classifiers, but TNS (TNSP and TNSU) performed constantly well for all criteria, followed by CFS. The worst performance is consistently found from CNS for all classifiers.

The shortest mean computational time was detected from subset evaluators, especially CFS, followed by CNS, but the filter ranking methods, TNSP and TNSU for NB were recorded to have the third and fourth best performance, followed by IG, TNS using C4.5, RLF and WRP. However, there was insufficient evidence to show a difference of mean processing time among methods (α=0.05).

In the future, it would be interesting to use different tree induction learning schemes and improved decision tree algorithms, e.g., a lookahead algorithm for ID3 (Esmeir and Markovitch 2004) that predicts the profitability of a split at a node by estimating its effect on deeper descendants of the node, to compare against TNS. It may be worthwhile to investigate which attributes are differently or commonly selected by different AS methods; this investigation was carried out in Chapter 3 for the Weed Risk Assessment (WRA) model and sea container contamination pathway, to understand which questions or items are important. Currently, the TNS software is a script that runs WEKA externally to generate the decision

tree and test attribute subsets, thus, it will be developed into a standalone system so that it can be more easily used and distributed.

2.5. References

Asuncion A, Newman DJ (2007) UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science. Available via http://mlearn.ics.uci.edu/MLRepository.html. Accessed on 28 July, 2008.

Baase S, Van Gelder A (2000) Computer algorithms; introduction to design& analysis, Addison-Wesley, Reading, 3rd ed.

Esmeir S, Markovitch *S* (2004) Lookahead-based algorithms for anytime induction of decision trees. In Proc. of 21st ICML 2004, 257-264.

Freitas AA (2002) Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, Berlin.

Fukuda K, Martin B (in press) Decision Trees as Information Source for Attribute Selection, In Proc. of the 2009 IEEE CIDM, 0-8.

Hall MA (1998) Correlation-Base Feature Selection for Machine Learning", PhD thesis, Dept of Computer Science, University of Waikato, Hamilton, 1998.

Hall MA (2000) Correlation-Based Feature Selection for discrete and numeric class machine learning. In Proc. of 17th ICML 2000, 359-366.

Hall MA, Holmes G (2003) Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Trans Knowl Eng 15: 1437-1447.

Jensen R, Shen Q (2007) Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst 15: 73-89.

Kira K, Rendell L (1992) A practical approach to feature selection. In Proc. 9th ICML 1992, 249-256.

Kononenko I (1994) Estimating attributes: Analysis and Extensions of Relief, In Proc. of 7th ECML, 171-182.

Konnenko I, Šimec E, Robnik-Šikonja M (1997) Overcoming the myopia of inductive learning algorithms with ReliefF, Appl Intell 7: 39-55.

Last M, Kandel A, Maimon, O (2001) Information-theoretic algorithm for feature selection. Pattern Recognit lett 22: 799-811.

Liu H, Setiono R (1996) A probabilistic approach to feature selection: A filter solution, In Proc. of 13th ICML 1996, 319-327.

McClave JT, Sincich T (2003) Annotated Instructor's Edition Statistics, Prentice Hall, New Jersey, 9th edn.

Minitab (2008) Help with statistics. Available on http://www.minitab.com/. Access on 28 July 2008.

Mitchell TM (1997) Machine learning, McGraw-Hill, New York.

Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data, Norwell, MA, Kluwer.

Quinlan JR (1986) Induction of decision trees. Mach Learn 1: 81-106.

Quinlan JR (1987) Rule induction with statistical data – a comparison with multiple regression. J Ope Res Soc 38; 347-352.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Schuschel D, Hsu C-N (1998) A weight analysis-based wrapper approach to neural nets feature subset selection, Tools with Artificial Intelligence, In Proc. of 10th IEEE ICTAI, 89 – 96.

Spate JM, Gibert K, Sànchez-Marrè M, Frank E, Comas J, Athanasiadis I, Letcher R (2006) Data mining as a tool for environmental scientists. In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-22, Burlington.

Wang D, Jiang L (2007) An improved attribute selection measure for decision tree induction, In Proc. 4th FSKD 2007, 110-117.

Wang J-F, Wang X-Z, Ha M-H (2002) Attribute selection's impact on robust of decision trees, In Proc. 1st ICML C, 1829-1832.

WEKA (2008) Documentation in Weka 3: Data Mining Software in Java, available via http://www.cs.waikato.ac.nz/ml/weka/. Accessed on 28 July 2008.

Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2nd edn.

2.6. Appendices

Appendix 2-1 Mean and standard deviation of classification accuracy over 10-fold cross validation (n = 10) using C4.5 with pruning.

Data set	C4.5-P	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	86.07 ± 0.42	86.02 ± 0.48	85.99 ± 0.44	86.07 ± 0.53	85.89 ± 0.55	86.08 ± 0.49	86.10 ± 0.45	85.94 ± 0.50
anneal	93.32 ± 2.73	92.98 ± 1.90	92.42 ± 2.16	92.31 ± 3.25	89.42 ± 2.54	93.20 ± 2.38	92.65 ± 2.53	86.63 ± 2.76
arrhythmia	64.36 ± 6.07	68.14 ± 5.69	67.90 ± 4.46	68.80 ± 4.73	66.39 ± 6.41	68.35 ± 4.13	68.35 ± 3.41	64.86 ± 6.87
audiology	77.09 ± 9.65	78.40 ± 8.45	78.83 ± 8.80	76.21 ± 6.92	77.57 ± 10.17	78.83 ± 8.05	78.83 ± 8.05	73.10 ± 9.54
balance-scale	78.06 ± 4.79	78.06 ± 4.79	78.06 ± 4.79	78.06 ± 4.79	74.09 ± 3.79	78.06 ± 4.79	78.06 ± 4.79	76.79 ± 4.18
bands	69.88 ± 3.89	71.74 ± 2.32	69.88 ± 3.89	76.00 ± 4.37	69.88 ± 3.89	69.88 ± 3.89	74.15 ± 4.05	57.81 ± 0.68
breast-cancer2	75.92 ± 7.64	75.23 ± 8.28	71.39 ± 9.78	71.72 ± 5.30	73.49 ± 8.03	72.80 ± 9.34	70.69 ± 7.29	72.09 ± 8.30
credit-screening	85.88 ± 5.18	86.47 ± 3.70	85.87 ± 5.46	85.73 ± 6.28	85.74 ± 5.60	86.31 ± 5.96	85.58 ± 5.21	86.02 ± 4.74
ecoli	83.69 ± 6.92	82.49 ± 6.54	82.78 ± 6.45	82.82 ± 7.61	83.69 ± 6.92	83.69 ± 6.92	83.69 ± 6.92	83.69 ± 6.92
flags	56.13 ± 7.25	57.66 ± 9.03	57.13 ± 8.83	64.95 ± 12.53	60.24 ± 8.39	58.16 ± 8.97	60.26 ± 7.56	59.66 ± 11.19
german credit	70.70 ± 4.45	71.60 ± 4.55	71.50 ± 3.37	71.60 ± 4.43	72.90 ± 3.11	73.50 ± 2.12	72.00 ± 2.75	73.60 ± 3.10
hepatitis2	80.00 ± 8.24	80.71 ± 8.80	83.21 ± 6.22	83.83 ± 7.57	81.29 ± 9.19	81.92 ± 4.06	82.58 ± 8.95	82.54 ± 8.08
horse-coli	84.33 ± 6.30	85.33 ± 5.26	85.67 ± 6.68	84.33 ± 6.68	84.33 ± 6.30	84.67 ± 6.32	86.33 ± 6.75	84.67 ± 5.92
hypothyroid	99.58 ± 0.33	99.66 ± 0.31	99.66 ± 0.31	99.63 ± 0.33	97.45 ± 0.80	99.58 ± 0.33	99.58 ± 0.33	99.60 ± 0.26
ionophere	91.75 ± 4.92	90.33 ± 6.60	90.04 ± 6.87	93.45 ± 4.04	91.75 ± 3.89	92.04 ± 3.69	91.19 ± 3.80	90.04 ± 4.48
iris	94.00 ± 6.63	94.00 ± 6.63	94.00 ± 6.63	91.33 ± 6.32	94.00 ± 6.63	92.00 ± 6.89	93.33 ± 6.29	92.00 ± 6.89
kr-vs-kp	99.34 ± 0.54	99.34 ± 0.45	99.34 ± 0.45	94.34 ± 1.29	94.05 ± 1.25	99.41 ± 0.45	99.41 ± 0.45	94.34 ± 1.29
labor-relations	85.00 ± 17.48	85.00 ± 17.48	82.50 ± 16.87	72.50 ± 18.45	85.00 ± 17.48	82.50 ± 16.87	82.50 ± 16.87	82.50 ± 16.87
letter recognition	88.06 ± 0.77	88.54 ± 0.58	88.56 ± 0.57	88.70 ± 0.55	88.27 ± 0.37	88.43 ± 0.71	88.40 ± 0.67	88.39 ± 0.75
mammographic	81.90 ± 4.80	82.31 ± 5.56	81.69 ± 4.83	81.69 ± 4.70	82.63 ± 4.43	82.42 ± 5.24	82.83 ± 4.58	81.90 ± 4.80
mushroom	100 ± 0.00	100 ± 0.00	100 ± 0.00	100 ± 0.00	99.02 ± 0.23	100 ± 0.00	100 ± 0.00	99.94 ± 0.13
pimadiabetes	74.34 ± 4.24	73.69 ± 2.96	73.82 ± 2.81	73.82 ± 3.52	75.00 ± 3.20	74.35 ± 3.65	74.08 ± 2.75	74.74 ± 3.35
primary-tumor	40.08 ± 8.77	40.11 ± 8.16	38.03 ± 6.29	40.12 ± 4.39	41.57 ± 6.12	42.19 ± 7.27	41.90 ± 7.10	39.49 ± 8.21
segment	93.58 ± 2.52	94.81 ± 2.65	94.81 ± 2.65	94.94 ± 2.63	94.69 ± 2.40	94.32 ± 2.74	94.69 ± 2.85	93.21 ± 3.74
sonar	74.50 ± 5.14	74.00 ± 6.60	74.00 ± 6.60	69.71 ± 5.52	76.45 ± 11.34	79.83 ± 5.31	73.62 ± 6.61	76.00 ± 7.69
soybean	89.76 ± 3.88	89.17 ± 3.72	89.76 ± 4.57	90.20 ± 3.50	88.87 ± 4.16	90.35 ± 3.57	89.76 ± 4.00	81.12 ± 3.84
splice	93.86 ± 0.88	94.04 ± 0.90	93.67 ± 0.83	94.01 ± 1.22	94.33 ± 1.17	93.76 ± 1.14	93.95 ± 1.19	51.88 ± 0.17
heart-c	79.52 ± 8.03	78.87 ± 5.49	77.54 ± 6.68	73.90 ± 3.81	74.22 ± 8.19	80.13 ± 8.49	81.17 ± 7.43	78.53 ± 7.92
waveform	76.04 ± 1.44	77.00 ± 0.96	76.84 ± 1.27	76.04 ± 1.18	76.82 ± 1.51	76.94 ± 1.83	76.88 ± 1.43	75.66 ± 1.45
vehicle	72.57 ± 3.31	72.58 ± 3.58	71.63 ± 3.57	68.91 ± 6.43	68.92 ± 3.62	73.17 ± 4.38	73.04 ± 5.10	73.53 ± 3.76
voting	96.33 ± 3.08	96.33 ± 3.08	96.09 ± 3.25	95.63 ± 3.65	95.64 ± 3.47	96.11 ± 3.39	95.63 ± 2.73	96.56 ± 2.88
wine	93.20 ± 3.70	93.79 ± 4.94	93.79 ± 4.94	92.09 ± 6.85	93.20 ± 3.70	94.31 ± 5.40	93.20 ± 4.54	92.71 ± 4.64
zoo	92.00 ± 7.89	89.09 ± 5.70	89.09 ± 5.70	92.09 ± 6.30	91.09 ± 5.67	91.09 ± 5.67	92.00 ± 6.32	91.09 ± 9.94
overall	82.45 ± 13.09	82.65 ± 12.74	82.29 ± 13.13	81.99 ± 12.58	82.06 ± 12.35	82.98 ± 12.44	82.92 ± 12.41	80.02 ± 13.74

Appendix 2-2 Mean and standard deviation classification accuracy over 10-fold cross validation (n = 10) using C4.5 without pruning.

Data set	C4.5-U	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	84.76 ± 0.34	85.81 ± 0.53	85.87 ± 0.55	85.72 ± 0.44	85.73 ± 0.51	85.27 ± 0.43	84.92 ± 0.40	84.94 ± 0.44
anneal	94.76 ± 1.98	94.65 ± 1.26	93.99 ± 1.84	92.87 ± 3.57	89.31 ± 2.86	94.88 ± 1.59	94.65 ± 1.96	90.65 ± 2.29
arrhythmia	62.81 ± 6.97	64.81 ± 5.58	63.70 ± 4.82	64.38 ± 5.28	64.61 ± 5.56	67.91 ± 4.14	64.36 ± 6.61	61.98 ± 6.20
audiology	77.11 ± 10.88	77.08 ± 9.26	77.09 ± 8.00	74.88 ± 8.27	76.23 ± 10.75	76.19 ± 10.46	76.64 ± 10.17	70.91 ± 12.58
balance-scale	79.02 ± 4.17	79.02 ± 4.17	79.02 ± 4.17	79.02 ± 4.17	73.94 ± 4.54	79.02 ± 4.17	79.02 ± 4.17	77.75 ± 3.82
bands	70.43 ± 7.83	71.56 ± 2.00	68.38 ± 7.20	76.56 ± 4.16	69.87 ± 7.98	70.43 ± 7.83	72.47 ± 3.74	60.39 ± 6.70
breast-cancer2	69.27 ± 9.76	75.23 ± 8.28	69.58 ± 5.65	71.72 ± 5.30	69.27 ± 6.06	71.40 ± 8.24	71.06 ± 7.90	67.86 ± 7.16
credit-screening	82.96 ± 5.11	83.69 ± 5.47	86.17 ± 6.10	85.73 ± 6.28	81.79 ± 4.75	84.85 ± 7.31	84.86 ± 5.75	83.25 ± 4.82
ecoli	82.50 ± 6.36	83.40 ± 7.05	83.40 ± 7.05	82.82 ± 7.61	82.50 ± 6.36	83.10 ± 6.88	83.10 ± 6.88	82.50 ± 6.36
flags	55.11 ± 7.63	59.74 ± 9.24	59.21 ± 9.51	63.84 ± 10.00	61.84 ± 8.59	60.84 ± 7.83	62.79 ± 7.27	57.63 ± 10.29
german credit	67.10 ± 4.70	70.70 ± 2.79	70.90 ± 4.48	70.40 ± 3.13	71.50 ± 3.03	72.20 ± 2.97	68.80 ± 5.07	69.90 ± 3.45
hepatitis2	80.67 ± 7.82	76.75 ± 10.58	81.25 ± 5.59	83.83 ± 7.57	78.08 ± 8.60	83.17 ± 6.90	79.38 ± 6.45	74.75 ± 9.84
horse-coli	83.00 ± 6.75	84.67 ± 7.06	85.00 ± 6.71	85.00 ± 6.71	83.67 ± 6.93	81.67 ± 6.14	86.00 ± 5.62	82.67 ± 4.92
hypothyroid	99.55 ± 0.31	99.63 ± 0.31	99.58 ± 0.38	99.58 ± 0.36	97.30 ± 0.81	99.58 ± 0.33	99.55 ± 0.31	99.44 ± 0.36
ionophere	91.75 ± 4.92	89.75 ± 6.17	89.47 ± 6.15	93.45 ± 4.04	91.46 ± 4.24	92.04 ± 3.69	91.19 ± 3.80	89.75 ± 4.47
iris	94.00 ± 6.63	94.00 ± 6.63	94.00 ± 6.63	91.33 ± 6.32	94.00 ± 6.63	92.00 ± 6.89	93.33 ± 6.29	92.00 ± 6.89
kr-vs-kp	99.31 ± 0.55	99.37 ± 0.47	99.28 ± 0.59	94.34 ± 1.29	94.18 ± 1.21	99.34 ± 0.48	99.31 ± 0.46	94.34 ± 1.29
labor-relations	82.50 ± 16.87	90.00 ± 17.48	87.50 ± 17.68	70.00 ± 19.72	85.00 ± 17.48	90.00 ± 17.48	85.00 ± 17.48	82.50 ± 16.87
letter recognition	88.09 ± 0.67	88.41 ± 0.61	88.45 ± 0.61	88.72 ± 0.54	88.27 ± 0.43	88.51 ± 0.66	88.46 ± 0.62	88.44 ± 0.67
mammographic	81.06 ± 5.08	82.83 ± 5.70	82.31 ± 5.54	82.00 ± 4.90	82.42 ± 4.27	83.35 ± 4.40	82.73 ± 3.29	81.06 ± 5.08
mushroom	100 ± 0.00	100 ± 0.00	100 ± 0.00	100 ± 0.00	99.02 ± 0.23	100 ± 0.00	100 ± 0.00	99.94 ± 0.13
pimadiabetes	74.08 ± 4.21	73.57 ± 2.67	73.31 ± 3.13	73.69 ± 3.39	74.61 ± 2.99	73.44 ± 2.88	73.95 ± 2.91	74.22 ± 3.20
primary-tumor	40.37 ± 9.07	41.59 ± 4.62	39.80 ± 6.12	39.81 ± 4.95	40.09 ± 6.53	40.14 ± 5.45	43.38 ± 5.85	40.96 ± 9.52
segment	93.46 ± 2.61	94.94 ± 2.70	94.94 ± 2.70	94.81 ± 2.65	94.69 ± 2.33	94.57 ± 2.74	94.69 ± 3.29	93.21 ± 3.74
sonar	74.50 ± 5.14	73.50 ± 7.10	73.50 ± 7.10	69.21 ± 5.20	76.45 ± 11.34	79.83 ± 4.82	73.62 ± 6.61	76.00 ± 7.69
soybean	89.02 ± 4.48	88.15 ± 3.39	88.59 ± 4.55	90.05 ± 3.04	89.02 ± 4.67	89.18 ± 3.95	89.91 ± 3.85	82.15 ± 5.41
splice	92.38 ± 1.48	93.07 ± 1.56	93.20 ± 1.66	92.98 ± 1.33	93.17 ± 1.07	93.51 ± 1.39	93.70 ± 1.04	52.63 ± 0.27
heart-c	79.20 ± 9.90	77.53 ± 6.33	76.89 ± 6.44	73.57 ± 4.31	76.23 ± 8.15	81.81 ± 7.06	80.49 ± 10.67	80.86 ± 10.06
waveform	76.04 ± 1.40	77.04 ± 0.96	76.76 ± 1.40	75.72 ± 1.12	76.76 ± 1.44	77.14 ± 1.72	76.84 ± 1.53	75.44 ± 1.50
vehicle	72.93 ± 3.19	73.53 ± 3.81	73.30 ± 4.15	68.91 ± 6.37	68.56 ± 3.11	73.17 ± 4.42	72.33 ± 4.12	73.53 ± 3.80
voting	96.10 ± 3.05	95.86 ± 3.21	95.63 ± 3.14	95.63 ± 3.65	94.94 ± 4.17	95.17 ± 2.93	95.18 ± 3.64	96.34 ± 3.06
wine	93.76 ± 3.39	93.24 ± 5.79	93.24 ± 5.79	91.54 ± 6.86	93.76 ± 3.39	93.20 ± 7.41	93.76 ± 4.28	92.71 ± 4.64
zoo	93.00 ± 6.75	90.00 ± 6.67	90.00 ± 6.67	92.09 ± 6.30	91.09 ± 5.67	91.09 ± 5.67	92.00 ± 6.32	91.09 ± 9.94
overall	81.84 ± 13.40	82.52 ± 12.70	82.22 ± 13.12	81.64 ± 12.92	81.50 ± 12.58	82.97 ± 12.61	82.65 ± 12.51	79.45 ± 13.85

Appendix 2-3 Mean and standard deviation of classification accuracy over 10-fold cross validation (n = 10) using naïve Bayes.

Data set	NB	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	83.23 ±0.62	83.22 ±0.60	83.26 ±0.60	80.67 ±1.65	79.95 ±0.41	83.84 ±0.67	83.97 ±0.64	82.30 ±0.61
anneal	64.58 ±4.20	80.39 ±3.97	80.73 ±3.72	68.84 ±14.50	60.36 ±5.48	78.84 ±3.45	89.42 ±3.14	60.24 ±1.89
arrhythmia	62.17 ±4.88	65.25 ±3.75	63.50 ±6.10	67.25 ±6.46	69.23 ±5.36	64.38 ±4.85	65.02 ±5.02	66.15 ±5.40
audiology	72.67 ±9.19	74.39 ±7.61	74.45 ±8.71	75.28 ±5.70	72.17 ±7.92	73.10 ±9.89	72.67 ±8.96	69.13 ±10.86
balance-scale	90.88 ±1.68	90.88 ±1.68	90.88 ±1.68	90.88 ±1.68	82.44 ±9.33	90.88 ±1.68	90.88 ±1.68	89.77 ±4.48
bands	72.08 ±7.48	69.32 ±4.14	70.05 ±6.97	69.12 ±7.16	70.05 ±7.01	71.90 ±5.58	74.13 ±7.49	67.46 ±6.10
breast-cancer2	72.39 ±7.70	74.89 ±8.32	70.31 ±8.75	70.68 ±6.38	73.76 ±6.88	73.46 ±8.34	67.84 ±5.26	72.40 ±5.43
credit-screening	78.31 ±5.63	84.72 ±6.14	86.02 ±5.50	85.73 ±6.28	75.40 ±6.43	85.73 ±6.28	86.03 ±5.17	78.45 ±5.44
ecoli	86.00 ±2.92	86.02 ±2.81	86.60 ±2.57	84.84 ±4.45	86.00 ±2.92	86.00 ±2.92	86.00 ±2.92	86.00 ±2.92
flags	44.89 ±6.64	51.97 ±11.10	52.05 ±10.62	51.11 ±10.80	53.97 ±13.31	46.92 ±8.53	51.45 ±11.08	46.29 ±13.82
german credit	74.30 ±4.60	74.60 ±3.69	75.60 ±3.78	72.40 ±4.62	73.30 ±4.00	75.10 ±5.00	74.20 ±2.57	73.30 ±3.59
hepatitis2	83.25 ±8.03	81.21 ±11.73	81.29 ±11.94	83.21 ±6.88	82.54 ±7.60	83.79 ±10.03	79.92 ±10.05	83.17 ±7.58
horse-coli	77.00 ±6.93	81.67 ±3.60	79.33 ±4.10	82.33 ±4.98	80.33 ±4.57	82.00 ±5.71	81.67 ±5.50	79.33 ±6.05
hypothyroid	95.33 ±0.79	95.25 ±0.64	95.28 ±0.69	95.18 ±0.73	94.62 ±0.79	95.36 ±0.69	95.36 ±0.75	95.18 ±0.72
ionophere	83.21 ±4.04	91.18 ±4.08	91.18 ±4.08	89.46 ±4.26	89.17 ±3.24	90.04 ±3.58	89.45 ±1.96	87.48 ±5.67
iris	95.33 ±3.22	95.33 ±4.50	95.33 ±4.50	92.00 ±5.26	95.33 ±4.50	94.00 ±4.92	95.33 ±4.50	92.67 ±5.84
kr-vs-kp	88.08 ±1.80	94.24 ±1.24	94.34 ±1.06	94.34 ±1.29	92.40 ±1.33	90.33 ±1.04	94.40 ±0.94	94.34 ±1.29
labor-relations	92.50 ±12.08	90.00 ±17.48	90.00 ±17.48	80.00 ±15.81	95.00 ±10.54	90.00 ±12.91	85.00 ±17.48	80.00 ±25.82
letter recognition	64.05 ±0.96	66.01 ±1.04	66.01 ±1.04	64.35 ±1.36	65.52 ±0.99	66.01 ±1.04	66.01 ±1.04	65.11 ±1.21
mammographic	82.62 ±3.57	80.85 ±4.79	82.21 ±4.02	80.44 ±4.50	82.00 ±4.20	82.73 ±3.51	82.52 ±3.63	82.62 ±3.57
mushroom	95.83 ±0.56	99.21 ±0.19	99.21 ±0.19	98.41 ±0.63	98.52 ±0.33	98.86 ±0.30	98.86 ±0.37	98.71 ±0.53
pimadiabetes	76.30 ±4.81	75.65 ±3.70	75.65 ±3.44	74.22 ±5.95	76.69 ±3.84	75.78 ±4.91	76.17 ±3.73	75.51 ±5.19
primary-tumor	49.57 ±7.41	48.38 ±6.67	49.57 ±7.67	41.88 ±4.68	48.09 ±5.46	49.27 ±6.98	49.56 ±7.47	49.28 ±6.49
segment	86.05 ±2.26	88.52 ±2.73	88.52 ±2.73	83.46 ±4.59	85.43 ±2.96	87.53 ±2.43	87.65 ±2.97	73.83 ±3.58
sonar	67.86 ±9.70	67.90 ±9.55	67.90 ±9.55	64.52 ±10.40	65.95 ±9.22	68.86 ±10.11	72.19 ±6.78	68.88 ±11.55
soybean	91.21 ±2.86	90.05 ±3.81	90.63 ±3.96	88.87 ±3.16	87.56 ±4.91	90.92 ±3.15	90.78 ±3.37	81.41 ±5.07
splice	95.33 ±1.13	95.30 ±1.25	96.18 ±1.04	94.36 ±1.38	95.86 ±1.07	95.80 ±1.02	96.33 ±1.11	52.63 ±0.27
heart-c	83.45 ±6.18	80.14 ±9.38	83.13 ±8.14	74.56 ±9.37	84.81 ±5.50	82.47 ±5.53	80.16 ±6.37	83.45 ±5.55
waveform	80.88 ±1.45	81.38 ±1.23	81.18 ±1.06	80.26 ±1.50	80.90 ±1.41	81.02 ±1.35	80.42 ±1.45	81.30 ±1.30
vehicle	44.89 ±6.53	52.81 ±6.74	52.93 ±6.66	45.49 ±5.34	45.85 ±3.91	44.06 ±6.38	46.66 ±6.17	46.43 ±6.08
voting	90.34 ±4.41	95.40 ±3.41	95.64 ±3.47	95.63 ±3.65	95.63 ±3.47	95.63 ±3.65	95.40 ±3.73	92.20 ±6.56
wine	97.75 ±2.91	93.17 ±5.99	93.17 ±5.99	93.76 ±5.73	97.19 ±3.96	97.16 ±3.00	97.16 ±3.00	95.46 ±4.51
zoo	97.00 ±4.83	87.09 ±4.90	88.09 ±4.27	92.00 ±9.19	93.09 ±6.71	94.09 ±6.94	97.00 ±4.83	94.00 ±8.43
overall	79.37 ±14.60	80.80 ±13.24	80.92 ±13.36	78.95 ±14.40	79.67 ±14.17	80.78 ±14.32	81.20 ±14.02	77.11 ±14.58

Appendix 2-4 Mean and standard deviation of relative reduction of attributes over 10-fold cross validation (n = 10) using C4.5 with pruning.

Data set	C4.5-P	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	15	10.1 ± 2.2	11.3 ±2.1	8.0 ± 2.1	5.0 ±0.0	11.3 ±1.1	13.3 ±0.5	12.0 ±0.0
anneal	37	8.8 ± 0.9	10.4 ±1.3	9.7 ± 2.6	5.0 ±0.0	17.5 ±5.3	13.8 ±2.3	7.9 ±0.3
arrhythmia	280	14.4 ± 9.2	14.6 ±9.9	8.5 ± 2.5	21.9 ±2.1	57.2 ±41.6	27.2 ±12.5	21.4 ±0.8
audiology	70	13.6 ± 0.8	16.2 ±1.5	9.1 ± 2.0	15.1 ±0.7	19.2 ±5.2	18.6 ±5.3	12.5 ±0.7
balance-scale	5	4.0 ± 0.0	4.0 ±0.0	4.0 ± 0.0	3.3 ±0.7	4.0 ±0.0	4.0 ±0.0	3.9 ±0.3
bands	40	1.3 ± 0.5	1.6 ±0.5	7.2 ± 1.7	2.0 ±0.0	5.2 ±0.4	13.7 ±2.5	6.9 ±0.9
breast-cancer2	10	2.1 ± 0.3	4.7 ±2.4	2.3 ± 0.9	3.9 ±0.9	4.8 ±2.5	5.9 ±3.2	7.8 ±0.4
credit-screening	16	5.7 ± 2.2	5.9 ±3.7	1.2 ± 0.6	6.5 ±0.5	8.6 ±4.2	8.5 ±4.4	12.6 ±0.7
ecoli	8	5.1 ± 0.6	5.2 ±0.6	4.7 ± 0.9	6.0 ±0.0	5.9 ±0.3	5.9 ±0.3	6.0 ±0.0
flags	27	6.3 ± 3.0	6.2 ±4.0	5.1 ± 1.7	4.9 ±0.3	7.7 ±7.8	9.1 ±6.9	7.8 ±0.4
german credit	21	8.6 ± 4.6	7.7 ±5.1	3.8 ± 1.9	4.0 ±0.7	5.8 ±1.8	9.1 ±6.3	12.3 ±0.7
hepatitis2	20	2.6 ± 2.1	2.5 ±2.7	1.9 ± 0.6	8.7 ±1.4	1.8 ±1.0	4.4 ±1.3	8.4 ±0.7
horse-coli	23	4.1 ± 1.4	4.2 ±2.3	4.2 ± 1.2	4.1 ±0.3	9.2 ±1.0	5.1 ±1.7	12.6 ±0.7
hypothyroid	30	6.6 ± 0.5	6.4 ±0.8	6.9 ± 0.3	5.9 ±0.6	16.5 ±1.6	21.8 ±0.4	8.4 ±0.8
ionophere	35	6.4 ± 2.4	6.4 ±2.4	5.0 ± 1.3	13.0 ±0.7	12.2 ±8.8	11.0 ±10.5	7.5 ±0.5
iris	5	1.2 ± 0.4	1.2 ±0.4	1.1 ± 0.3	2.0 ±0.0	1.6 ±1.0	1.5 ±1.0	1.0 ±0.0
kr-vs-kp	37	21.8 ± 0.6	24.1 ±1.6	5.0 ± 0.0	6.8 ±0.4	34.9 ±0.7	29.0 ±2.9	5.9 ±0.3
labor-relations	17	1.2 ± 0.4	1.4 ±1.0	1.6 ± 0.8	6.2 ±0.8	1.7 ±1.3	1.8 ±1.8	3.7 ±0.5
letter recognition	17	11.0 ± 0.9	10.9 ±0.9	9.9 ± 0.7	10.6 ±0.5	12.0 ±1.2	11.0 ±1.2	12.6 ±0.7
mammographic	6	2.8 ± 0.9	3.6 ±1.2	2.6 ± 1.1	3.5 ±0.5	3.1 ±0.7	4.5 ±0.7	5.0 ±0.0
mushroom	23	5.0 ± 0.0	5.0 ±0.0	4.9 ± 0.3	4.0 ±0.0	9.0 ±0.0	8.0 ±0.0	4.9 ±0.3
pimadiabetes	9	3.4 ± 1.1	3.4 ±1.1	3.6 ± 1.0	4.2 ±0.4	4.0 ±2.2	3.4 ±1.7	6.0 ±0.0
primary-tumor	18	9.0 ± 4.4	10.4 ±3.8	6.6 ± 1.4	11.9 ±0.9	9.8 ±3.3	12.1 ±3.3	15.7 ±0.5
segment	20	5.0 ± 2.3	5.0 ±2.3	6.3 ± 1.3	7.1 ±0.3	11.2 ±0.6	10.2 ±1.7	8.0 ±0.5
sonar	61	8.2 ± 2.6	8.2 ±2.6	5.2 ± 2.9	17.5 ±1.3	20.1 ±5.2	20.9 ±11.4	12.8 ±0.8
soybean	35	19.6 ± 4.2	22.6 ±4.2	15.2 ± 1.8	21.5 ±1.1	29.5 ±1.4	30.0 ±3.0	14.6 ±0.7
splice	62	10.5 ± 2.5	14.5 ±7.7	10.4 ± 1.6	21.4 ±1.4	9.0 ±2.3	17.9 ±13.9	1.0 ±0.0
heart-c	14	3.6 ± 0.7	4.2 ±1.8	3.4 ± 2.2	6.6 ±1.0	3.5 ±0.7	5.1 ±0.9	10.0 ±0.5
waveform	22	11.2 ± 3.4	10.6 ±2.5	9.8 ± 1.8	15.6 ±0.5	14.3 ±2.4	11.8 ±2.7	13.2 ±0.4
vehicle	19	13.6 ± 2.4	13.2 ±2.7	7.0 ± 2.6	8.9 ±1.4	16.6 ±1.3	15.8 ±1.0	16.6 ±0.5
voting	17	4.6 ± 0.7	5.8 ±1.0	1.0 ± 0.0	4.2 ±0.6	12.9 ±1.2	10.3 ±5.9	8.6 ±1.1
wine	14	3.2 ± 0.8	3.2 ±0.8	3.7 ± 0.7	10.8 ±0.4	3.7 ±1.1	5.3 ±1.9	4.8 ±0.4
zoo	17	6.0 ± 0.9	6.0 ±0.9	5.0 ± 0.5	8.9 ±1.7	10.4 ±1.3	9.6 ±1.5	5.3 ±0.5
overall	32±47	7.3 ±5.1	7.9 ±5.6	5.6 ±3.2	8.5 ±5.7	11.9 ±11.1	11.5 ±7.6	9.0 ±4.6

Appendix 2-5 Mean and standard deviation of relative reduction of attributes over 10-fold cross validation (n = 10) using C4.5 without pruning.

Data set	C4.5-U	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	15	5.2 ± 0.8	5.8 ± 0.4	8.0 ± 2.1	5.0 ± 0.0	9.4 ± 0.7	13.3 ± 0.5	12.0 ± 0.0
anneal	37	8.5 ± 1.2	9.2 ± 0.9	9.7 ± 2.6	5.0 ± 0.0	15.3 ± 6.0	13.6 ± 2.2	7.9 ± 0.3
arrhythmia	280	21.6 ± 7.7	27.6 ± 10.2	8.5 ± 2.5	21.9 ± 2.1	70.3 ± 49.6	36.5 ± 23.0	21.4 ± 0.8
audiology	70	13.7 ± 0.9	14.0 ± 3.1	9.1 ± 2.0	15.1 ± 0.7	18.5 ± 5.6	19.0 ± 5.2	12.5 ± 0.7
balance-scale	5	4.0 ± 0.0	4.0 ± 0.0	4.0 ± 0.0	3.3 ± 0.7	4.0 ± 0.0	4.0 ± 0.0	3.9 ± 0.3
bands	40	1.3 ± 0.5	1.8 ± 0.8	7.2 ± 1.7	2.0 ± 0.0	8.6 ± 5.4	11.7 ± 3.6	6.9 ± 0.9
breast-cancer2	10	2.2 ± 0.6	1.3 ± 0.9	2.3 ± 0.9	3.9 ± 0.9	2.6 ± 1.6	2.1 ± 1.0	7.8 ± 0.4
credit-screening	16	4.6 ± 2.3	2.4 ± 1.0	1.2 ± 0.6	6.5 ± 0.5	3.6 ± 2.0	3.7 ± 1.3	12.6 ± 0.7
ecoli	8	5.2 ± 0.6	5.3 ± 0.7	4.7 ± 0.9	6.0 ± 0.0	5.9 ± 0.3	5.9 ± 0.3	6.0 ± 0.0
flags	27	5.5 ± 2.9	5.8 ± 3.4	5.1 ± 1.7	4.9 ± 0.3	3.6 ± 3.7	3.4 ± 2.8	7.8 ± 0.4
german credit	21	5.5 ± 2.4	3.3 ± 1.1	3.8 ± 1.9	4.0 ± 0.7	3.2 ± 1.0	3.0 ± 1.1	12.3 ± 0.7
hepatitis2	20	4.9 ± 2.2	2.9 ± 2.5	1.9 ± 0.6	8.7 ± 1.4	1.8 ± 1.2	6.0 ± 4.2	8.4 ± 0.7
horse-coli	23	4.7 ± 1.6	5.9 ± 5.4	4.2 ± 1.2	4.1 ± 0.3	10.3 ± 4.5	4.6 ± 2.4	12.6 ± 0.7
hypothyroid	30	7.8 ± 0.6	7.7 ± 1.5	6.9 ± 0.3	5.9 ± 0.6	17.4 ± 1.3	21.8 ± 0.4	8.4 ± 0.8
ionophere	35	6.9 ± 2.3	7.6 ± 2.1	5.0 ± 1.3	13.0 ± 0.7	11.3 ± 8.9	11.3 ± 10.3	7.5 ± 0.5
iris	5	1.3 ± 0.5	1.3 ± 0.5	1.1 ± 0.3	2.0 ± 0.0	1.4 ± 0.5	1.3 ± 0.5	1.0 ± 0.0
kr-vs-kp	37	22.0 ± 0.7	25.5 ± 2.3	5.0 ± 0.0	6.8 ± 0.4	34.8 ± 0.6	33.4 ± 4.2	5.9 ± 0.3
labor-relations	17	1.6 ± 0.5	1.7 ± 0.8	1.6 ± 0.8	6.2 ± 0.8	1.8 ± 0.6	2.1 ± 1.4	3.7 ± 0.5
letter recognition	17	10.5 ± 0.8	10.4 ± 0.7	9.9 ± 0.7	10.6 ± 0.5	11.5 ± 1.0	10.7 ± 0.9	12.6 ± 0.7
mammographic	6	2.3 ± 0.7	3.0 ± 1.2	2.6 ± 1.1	3.5 ± 0.5	2.6 ± 0.5	4.4 ± 0.7	5.0 ± 0.0
mushroom	23	5.0 ± 0.0	5.0 ± 0.0	4.9 ± 0.3	4.0 ± 0.0	9.0 ± 0.0	8.0 ± 0.0	4.9 ± 0.3
pimadiabetes	9	3.8 ± 2.0	4.0 ± 2.1	3.6 ± 1.0	4.2 ± 0.4	3.4 ± 2.6	3.8 ± 2.0	6.0 ± 0.0
primary-tumor	18	8.4 ± 3.5	9.3 ± 3.6	6.6 ± 1.4	11.9 ± 0.9	9.6 ± 3.3	10.1 ± 2.7	15.7 ± 0.5
segment	20	4.6 ± 2.0	4.6 ± 2.0	6.3 ± 1.3	7.1 ± 0.3	11.5 ± 0.7	10.5 ± 2.1	8.0 ± 0.5
sonar	61	8.2 ± 2.6	8.2 ± 2.6	5.2 ± 2.9	17.5 ± 1.3	19.8 ± 3.9	20.8 ± 11.4	12.8 ± 0.8
soybean	35	18.3 ± 2.9	20.1 ± 4.5	15.2 ± 1.8	21.5 ± 1.1	29.2 ± 1.9	29.2 ± 1.3	14.6 ± 0.7
splice	62	10.3 ± 3.2	9.4 ± 7.0	10.4 ± 1.6	21.4 ± 1.4	8.0 ± 0.0	8.1 ± 2.9	1.0 ± 0.0
heart-c	14	3.3 ± 0.5	4.4 ± 2.1	3.4 ± 2.2	6.6 ± 1.0	3.3 ± 0.5	5.2 ± 1.6	10.0 ± 0.5
waveform	22	10.6 ± 2.6	10.9 ± 2.6	9.8 ± 1.8	15.6 ± 0.5	14.0 ± 2.1	11.6 ± 2.1	13.2 ± 0.4
vehicle	19	12.9 ± 2.3	12.8 ± 2.1	7.0 ± 2.6	8.9 ± 1.4	16.7 ± 1.1	15.2 ± 2.1	16.6 ± 0.5
voting	17	3.8 ± 0.9	4.9 ± 1.3	1.0 ± 0.0	4.2 ± 0.6	9.2 ± 6.0	5.4 ± 2.6	8.6 ± 1.1
wine	14	3.2 ± 0.8	3.2 ± 0.8	3.7 ± 0.7	10.8 ± 0.4	3.6 ± 1.1	5.4 ± 1.9	4.8 ± 0.4
zoo	17	6.3 ± 1.1	6.3 ± 1.1	5.0 ± 0.5	8.9 ± 1.7	10.2 ± 1.4	9.6 ± 1.5	5.3 ± 0.5
overall	32±47	7.2 ± 5.4	7.6 ± 6.4	5.6 ± 3.2	8.5 ± 5.7	11.7 ± 13.1	10.7 ± 9.0	9.0 ± 4.6

Appendix 2-6 Mean and standard deviation of relative reduction of attributes over 10-fold cross validation (n = 10) using naïve Bayes.

Data set	NB	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	15	13.3 ± 0.7	13.2 ± 0.4	8.0 ± 2.1	5.0 ± 0.0	8.0 ± 0.0	12.1 ± 0.3	12.0 ± 0.0
anneal	37	6.8 ± 2.4	8.1 ± 1.2	9.7 ± 2.6	5.0 ± 0.0	9.6 ± 1.0	6.9 ± 0.3	7.9 ± 0.3
arrhythmia	280	27.5 ± 7.2	27.8 ± 12.3	8.5 ± 2.5	21.9 ± 2.1	23.4 ± 18.6	28.5 ± 10.6	21.4 ± 0.8
audiology	70	13.6 ± 1.8	14.4 ± 3.5	9.1 ± 2.0	15.1 ± 0.7	30.9 ± 12.3	32.1 ± 13.0	12.5 ± 0.7
balance-scale	5	4.0 ± 0.0	4.0 ± 0.0	4.0 ± 0.0	3.3 ± 0.7	4.0 ± 0.0	4.0 ± 0.0	3.9 ± 0.3
bands	40	1.0 ± 0.0	2.0 ± 0.7	7.2 ± 1.7	2.0 ± 0.0	16.7 ± 14.9	13.1 ± 3.0	6.9 ± 0.9
breast-cancer2	10	2.3 ± 0.7	5.2 ± 2.9	2.3 ± 0.9	3.9 ± 0.9	3.0 ± 2.3	6.6 ± 3.2	7.8 ± 0.4
credit-screening	16	1.7 ± 1.2	2.7 ± 1.6	1.2 ± 0.6	6.5 ± 0.5	1.0 ± 0.0	6.9 ± 1.8	12.6 ± 0.7
ecoli	8	5.6 ± 0.5	5.7 ± 0.5	4.7 ± 0.9	6.0 ± 0.0	6.2 ± 0.4	6.2 ± 0.4	6.0 ± 0.0
flags	27	3.7 ± 2.1	4.9 ± 2.4	5.1 ± 1.7	4.9 ± 0.3	12.7 ± 9.3	13.0 ± 7.5	7.8 ± 0.4
german credit	21	12.0 ± 3.7	13.9 ± 2.7	3.8 ± 1.9	4.0 ± 0.7	16.0 ± 3.9	15.3 ± 3.7	12.3 ± 0.7
hepatitis2	20	4.0 ± 2.8	6.5 ± 4.0	1.9 ± 0.6	8.7 ± 1.4	9.3 ± 2.9	9.7 ± 6.0	8.4 ± 0.7
horse-coli	23	3.8 ± 1.0	5.1 ± 2.6	4.2 ± 1.2	4.1 ± 0.3	5.5 ± 1.4	4.2 ± 1.9	12.6 ± 0.7
hypothyroid	30	7.8 ± 0.9	8.4 ± 1.4	6.9 ± 0.3	5.9 ± 0.6	14.7 ± 4.4	22.7 ± 0.7	8.4 ± 0.8
ionophere	35	8.1 ± 2.2	8.4 ± 2.5	5.0 ± 1.3	13.0 ± 0.7	7.3 ± 2.8	6.7 ± 0.8	7.5 ± 0.5
iris	5	2.0 ± 0.0	2.0 ± 0.0	1.1 ± 0.3	2.0 ± 0.0	2.0 ± 0.5	2.2 ± 0.4	1.0 ± 0.0
kr-vs-kp	37	5.7 ± 1.5	8.2 ± 5.1	5.0 ± 0.0	6.8 ± 0.4	3.4 ± 1.3	8.4 ± 2.0	5.9 ± 0.3
labor-relations	17	1.9 ± 0.3	2.4 ± 0.5	1.6 ± 0.8	6.2 ± 0.8	6.6 ± 3.9	6.5 ± 1.8	3.7 ± 0.5
letter recognition	17	11.2 ± 0.6	11.2 ± 0.6	9.9 ± 0.7	10.6 ± 0.5	11.0 ± 0.0	11.0 ± 0.0	12.6 ± 0.7
mammographic	6	2.7 ± 1.1	3.7 ± 0.7	2.6 ± 1.1	3.5 ± 0.5	4.1 ± 0.3	4.9 ± 0.3	5.0 ± 0.0
mushroom	23	4.0 ± 0.0	4.0 ± 0.0	4.9 ± 0.3	4.0 ± 0.0	2.9 ± 0.3	2.6 ± 0.5	4.9 ± 0.3
pimadiabetes	9	3.9 ± 1.2	3.4 ± 1.3	3.6 ± 1.0	4.2 ± 0.4	3.2 ± 1.1	2.6 ± 1.3	6.0 ± 0.0
primary-tumor	18	13.7 ± 0.7	14.5 ± 1.0	6.6 ± 1.4	11.9 ± 0.9	13.4 ± 1.3	14.0 ± 1.4	15.7 ± 0.5
segment	20	10.9 ± 1.7	11.0 ± 1.8	6.3 ± 1.3	7.1 ± 0.3	13.9 ± 2.0	15.5 ± 1.1	8.0 ± 0.5
sonar	61	2.3 ± 2.5	2.3 ± 2.5	5.2 ± 2.9	17.5 ± 1.3	2.2 ± 1.8	2.6 ± 1.3	12.8 ± 0.8
soybean	35	20.8 ± 3.2	22.3 ± 2.4	15.2 ± 1.8	21.5 ± 1.1	32.0 ± 0.9	31.8 ± 2.3	14.6 ± 0.7
splice	62	14.3 ± 2.7	27.6 ± 4.3	10.4 ± 1.6	21.4 ± 1.4	29.1 ± 7.3	22.1 ± 2.6	1.0 ± 0.0
heart-c	14	6.1 ± 2.1	8.1 ± 2.1	3.4 ± 2.2	6.6 ± 1.0	6.2 ± 0.8	8.2 ± 2.7	10.0 ± 0.5
waveform	22	11.0 ± 2.6	11.8 ± 2.1	9.8 ± 1.8	15.6 ± 0.5	16.5 ± 0.8	14.0 ± 2.4	13.2 ± 0.4
vehicle	19	7.1 ± 2.3	7.8 ± 2.7	7.0 ± 2.6	8.9 ± 1.4	17.4 ± 1.0	10.0 ± 8.2	16.6 ± 0.5
voting	17	2.7 ± 0.5	3.0 ± 0.0	1.0 ± 0.0	4.2 ± 0.6	1.0 ± 0.0	1.4 ± 0.7	8.6 ± 1.1
wine	14	4.3 ± 1.3	4.6 ± 1.7	3.7 ± 0.7	10.8 ± 0.4	8.2 ± 3.3	9.4 ± 3.3	4.8 ± 0.4
zoo	17	6.4 ± 1.2	6.5 ± 1.3	5.0 ± 0.5	8.9 ± 1.7	10.6 ± 2.3	12.3 ± 1.3	5.3 ± 0.5
overall	32±47	7.5 ± 5.9	8.6 ± 6.7	5.6 ± 3.2	8.5 ± 5.7	10.7 ± 8.5	11.1 ± 8.2	9.0 ± 4.6

Appendix 2-7 Mean and standard deviation of relative reduction of decision tree size over 10-fold cross validation (n = 10) using C4.5 with pruning.

Data set	C4.5-P	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	588.7 ± 72.8	421.4 ± 157.1	379.7 ± 147.8	200.1 ± 193.7	73.6 ± 5.1	485.9 ± 135.2	491.3 ± 102.0	354.0 ± 39.1
anneal	32.8 ± 7.5	31.7 ± 3.1	27.6 ± 1.3	34.1 ± 3.1	30.6 ± 2.5	32.6 ± 7.6	29.5 ± 4.8	22.4 ± 2.1
arrhythmia	40.6 ± 1.8	29.0 ± 10.3	27.4 ± 7.1	26.3 ± 3.5	38.1 ± 3.2	36.9 ± 2.7	36.2 ± 2.8	43.4 ± 6.4
audiology	30.6 ± 2.2	27.8 ± 2.1	27.9 ± 1.9	24.1 ± 3.0	26.6 ± 2.6	29.5 ± 3.0	29.5 ± 2.8	28.7 ± 4.3
balance-scale	40.5 ± 5.1	40.5 ± 5.1	40.5 ± 5.1	40.5 ± 5.1	23.8 ± 16.3	40.5 ± 5.1	40.5 ± 5.1	37.6 ± 9.7
bands	2.8 ± 1.0	7.5 ± 9.0	2.0 ± 0.0	33.8 ± 8.0	2.0 ± 0.0	2.4 ± 0.8	107.5 ± 31.5	1.0 ± 0.0
breast-cancer2	4.8 ± 1.9	4.0 ± 0.0	5.1 ± 3.2	3.2 ± 1.5	10.4 ± 1.0	5.6 ± 2.9	3.9 ± 1.4	16.3 ± 5.7
credit-screening	22.6 ± 11.4	13.4 ± 10.1	14.0 ± 11.9	2.0 ± 0.0	15.9 ± 11.5	16.0 ± 12.6	24.8 ± 13.5	22.4 ± 10.7
ecoli	18.9 ± 2.5	19.4 ± 2.2	19.1 ± 2.6	19.3 ± 2.8	18.9 ± 2.5	19.2 ± 2.4	19.2 ± 2.4	18.9 ± 2.5
flags	27.0 ± 4.9	20.7 ± 4.6	19.2 ± 6.2	16.2 ± 3.6	25.7 ± 5.3	17.7 ± 7.9	19.0 ± 7.7	26.3 ± 6.2
german credit	89.3 ± 21.2	48.2 ± 25.2	43.7 ± 31.2	18.4 ± 13.1	19.3 ± 7.4	41.1 ± 12.2	50.5 ± 32.1	65.4 ± 23.8
hepatitis2	8.8 ± 2.6	3.8 ± 2.1	2.6 ± 1.1	3.6 ± 1.3	5.0 ± 1.6	3.5 ± 1.2	2.1 ± 0.3	5.6 ± 2.1
horse-coli	7.5 ± 2.1	5.1 ± 2.1	4.4 ± 1.6	4.7 ± 1.6	4.7 ± 1.3	8.1 ± 2.4	5.4 ± 1.5	6.3 ± 2.4
hypothyroid	14.6 ± 0.7	12.5 ± 0.8	12.0 ± 0.5	12.5 ± 0.5	7.7 ± 1.6	13.5 ± 1.4	14.6 ± 0.7	9.1 ± 0.9
ionophere	13.4 ± 2.2	10.0 ± 3.2	10.1 ± 3.0	7.2 ± 2.0	12.6 ± 1.9	8.6 ± 2.7	9.7 ± 3.5	11.6 ± 4.2
iris	4.5 ± 0.7	3.4 ± 0.8	3.4 ± 0.8	3.2 ± 0.6	4.5 ± 0.7	3.6 ± 0.8	3.5 ± 0.8	3.0 ± 0.0
kr-vs-kp	29.9 ± 1.7	29.0 ± 2.3	28.9 ± 2.4	6.0 ± 0.0	5.7 ± 0.5	30.2 ± 1.5	29.5 ± 1.9	6.0 ± 0.0
labor-relations	3.6 ± 1.0	2.4 ± 0.8	2.5 ± 1.1	2.7 ± 1.2	4.0 ± 0.8	2.7 ± 1.2	2.5 ± 1.1	2.1 ± 0.3
letter recognition	1165 ± 9.8	1155 ± 16.5	1156 ± 15.9	1160 ± 15.0	1159 ± 22.3	1153 ± 16.8	1157 ± 19.6	1163 ± 18.6
mammographic	6.3 ± 2.2	5.9 ± 2.6	5.9 ± 2.3	4.8 ± 1.5	5.8 ± 1.3	6.0 ± 1.9	6.8 ± 2.0	6.3 ± 2.2
mushroom	24.0 ± 0.0	24.0 ± 0.0	24.0 ± 0.0	30.9 ± 1.4	16.0 ± 0.0	29.0 ± 0.0	24.2 ± 4.6	31.9 ± 1.5
pimadiabetes	21.0 ± 8.7	9.0 ± 3.3	8.9 ± 3.3	9.6 ± 5.8	10.2 ± 2.3	10.8 ± 7.8	8.0 ± 5.1	14.6 ± 5.7
primary-tumor	46.0 ± 3.4	26.5 ± 14.2	32.4 ± 12.6	13.8 ± 4.2	36.5 ± 8.4	26.7 ± 12.3	33.1 ± 14.2	47.3 ± 4.5
segment	24.2 ± 2.4	20.8 ± 1.5	20.8 ± 1.5	20.7 ± 1.4	21.2 ± 2.0	22.7 ± 1.8	21.2 ± 2.1	26.7 ± 1.4
sonar	14.5 ± 1.7	13.1 ± 3.8	13.1 ± 3.8	7.4 ± 4.9	15.2 ± 1.6	15.0 ± 1.8	13.7 ± 3.1	15.3 ± 2.3
soybean	69.7 ± 9.9	58.6 ± 10.6	63.7 ± 7.5	54.4 ± 5.6	72.3 ± 5.7	59.7 ± 6.2	63.5 ± 7.9	73.7 ± 8.8
splice	165.3 ± 14.6	147.7 ± 17.8	144.4 ± 16.0	140.0 ± 8.1	163.9 ± 9.9	128.4 ± 10.6	153.9 ± 10.8	1.0 ± 0.0
heart-c	25.5 ± 6.5	12.7 ± 2.4	12.7 ± 3.1	10.1 ± 7.3	16.4 ± 3.1	13.1 ± 2.2	15.3 ± 2.1	20.9 ± 4.3
waveform	274.5 ± 18.4	198.1 ± 49.8	196.2 ± 49.0	185.2 ± 35.6	256.5 ± 18.9	238.4 ± 34.8	202.9 ± 38.3	239.6 ± 24.5
vehicle	64.1 ± 10.5	63.2 ± 9.7	62.2 ± 8.8	59.4 ± 13.6	58.7 ± 7.2	66.0 ± 12.0	61.7 ± 6.5	64.5 ± 11.4
voting	5.8 ± 0.4	5.6 ± 0.7	5.6 ± 0.5	2.0 ± 0.0	4.2 ± 0.9	5.6 ± 0.7	4.9 ± 1.3	5.7 ± 0.5
wine	5.7 ± 0.9	5.3 ± 0.9	5.3 ± 0.9	5.6 ± 0.5	5.7 ± 0.9	5.2 ± 0.4	5.7 ± 1.3	6.3 ± 0.7
zoo	8.3 ± 0.8	6.9 ± 0.7	6.8 ± 0.6	7.7 ± 0.5	7.6 ± 0.5	7.2 ± 0.4	7.0 ± 0.0	8.0 ± 0.5
overall	87.9 ± 222.3	75.2 ± 209.5	73.6 ± 207.7	65.7 ± 202.4	66.0 ± 202.6	78.3 ± 213.4	81.8 ± 213	72.9 ± 208.1

Appendix 2-8 Mean and standard deviation of relative reduction of decision tree size over 10-fold cross validation (n = 10) using C4.5 without pruning.

Data set	C4.5-U	TNSP	TNSU	WRP	CFS	IG	RLF	CNS
adult	8381.0 ± 206.0	203.1 ± 54.7	248.4 ± 42.8	1538 ± 2019	223.4 ± 16.7	3519 ± 726.8	8146 ± 262.5	6452 ± 206.5
anneal	64.1 ± 9.6	45.0 ± 8.2	51.3 ± 6.0	44.8 ± 6.5	37.9 ± 2.1	62.0 ± 11.5	61.5 ± 11.4	46.8 ± 2.4
arrhythmia	50.7 ± 2.1	50.5 ± 3.4	49.0 ± 3.7	52.7 ± 5.1	52.9 ± 4.0	49.6 ± 2.4	55.1 ± 5.3	56.1 ± 5.3
audiology	41.6 ± 5.5	35.4 ± 3.4	36.6 ± 7.2	31.9 ± 4.5	38.9 ± 3.7	49.0 ± 4.9	47.9 ± 5.7	45.6 ± 5.2
balance-scale	58.1 ± 1.4	58.1 ± 1.4	58.1 ± 1.4	58.1 ± 1.4	36.5 ± 19.1	58.1 ± 1.4	58.1 ± 1.4	55.0 ± 9.9
bands	442.9 ± 25.4	8.1 ± 9.7	261.6 ± 223.4	52.5 ± 13.3	347.6 ± 182.1	486.1 ± 135.8	283.9 ± 151.7	433.8 ± 0.4
breast-cancer2	106.4 ± 14.1	4.6 ± 1.9	7.5 ± 17.0	5.6 ± 5.7	39.0 ± 25.6	24.2 ± 25.8	14.1 ± 23.8	113.4 ± 15.9
credit-screening	112.5 ± 15.7	17.6 ± 12.7	11.8 ± 6.8	2.0 ± 0.0	63.4 ± 8.1	26.6 ± 18.5	61.5 ± 35.7	110.9 ± 15.1
ecoli	22.6 ± 2.5	22.3 ± 2.8	22.6 ± 2.5	22.1 ± 2.8	22.6 ± 2.5	22.9 ± 2.7	22.9 ± 2.7	22.6 ± 2.5
flags	38.1 ± 6.3	23.1 ± 6.6	23.1 ± 6.4	20.4 ± 4.0	42.2 ± 7.2	20.9 ± 12.9	18.7 ± 9.8	43.3 ± 4.6
german credit	285.8 ± 14.2	99.3 ± 86.5	49.9 ± 49.0	54.8 ± 71.8	44.8 ± 24.9	66.4 ± 73.3	82.2 ± 78.7	330.0 ± 26.5
hepatitis2	15.2 ± 1.5	8.4 ± 4.1	4.3 ± 3.0	4.1 ± 1.6	12.7 ± 2.6	4.5 ± 1.4	7.9 ± 4.6	13.1 ± 3.2
horse-coli	22.7 ± 2.4	11.5 ± 6.5	9.8 ± 8.7	7.8 ± 4.0	9.7 ± 2.8	20.3 ± 6.3	9.0 ± 5.4	19.8 ± 2.7
hypothyroid	16.9 ± 2.1	14.8 ± 0.9	15.4 ± 1.8	15.2 ± 2.0	10.5 ± 1.8	17.6 ± 2.8	16.9 ± 2.1	20.2 ± 5.3
ionophere	13.8 ± 2.3	11.2 ± 3.7	11.7 ± 2.9	7.7 ± 3.4	14.0 ± 1.9	9.7 ± 3.1	9.7 ± 3.5	12.7 ± 4.7
iris	4.9 ± 0.7	3.5 ± 0.8	3.5 ± 0.8	3.2 ± 0.6	4.9 ± 0.7	3.6 ± 0.8	3.5 ± 0.8	3.0 ± 0.0
kr-vs-kp	38.2 ± 3.0	32.4 ± 2.8	34.8 ± 2.3	6.0 ± 0.0	8.5 ± 1.0	35.8 ± 2.0	37.2 ± 3.1	6.0 ± 0.0
labor-relations	6.5 ± 1.3	3.4 ± 1.3	3.5 ± 1.7	2.8 ± 1.1	5.2 ± 1.5	3.6 ± 1.2	3.4 ± 1.5	2.6 ± 1.1
letter recognition	1278.3 ± 15.9	1298 ± 16.4	1300 ± 15.2	1305 ± 17.8	1299 ± 17.9	1289 ± 17.8	1294 ± 20.4	1279 ± 16.0
mammographic	23.4 ± 4.9	9.5 ± 5.4	12.9 ± 7.6	8.4 ± 3.6	17.2 ± 6.5	9.7 ± 2.2	19.2 ± 7.9	23.4 ± 4.9
mushroom	24.0 ± 0.0	24.0 ± 0.0	24.0 ± 0.0	30.9 ± 1.4	16.0 ± 0.0	29.0 ± 0.0	29.0 ± 2.1	31.9 ± 1.5
pimadiabetes	25.9 ± 7.3	12.4 ± 9.7	13.9 ± 10.3	11.4 ± 7.4	12.8 ± 3.9	10.4 ± 11.1	10.5 ± 7.3	20.0 ± 4.2
primary-tumor	66.6 ± 4.6	37.2 ± 24.4	43.3 ± 24.4	20.0 ± 8.8	57.6 ± 6.9	37.0 ± 20.0	42.2 ± 14.8	67.8 ± 5.7
segment	25.0 ± 2.6	21.6 ± 2.7	21.6 ± 2.7	22.5 ± 2.0	24.6 ± 1.6	25.1 ± 2.3	23.4 ± 2.0	27.4 ± 1.6
sonar	14.5 ± 1.7	13.4 ± 3.4	13.4 ± 3.4	7.9 ± 5.9	15.4 ± 1.6	15.5 ± 2.0	14.2 ± 3.4	15.3 ± 2.3
soybean	105.5 ± 6.8	81.6 ± 13.6	92.4 ± 11.3	66.8 ± 11.6	104.9 ± 12.9	94.4 ± 9.3	95.2 ± 5.5	144.3 ± 14.9
splice	3233.3 ± 1792	333.9 ± 88.4	605.1 ± 1030	382.2 ± 44.6	402.0 ± 16.9	7574 ± 1536	265.8 ± 49.3	3178 ± 0.0
heart-c	47.5 ± 3.0	13.2 ± 2.9	17.4 ± 10.6	14.8 ± 12.8	37.7 ± 5.8	14.4 ± 2.0	22.6 ± 14.3	45.6 ± 2.5
waveform	289.7 ± 15.3	215.6 ± 35.7	220.0 ± 40.3	206.4 ± 35.1	273.3 ± 17.6	256.9 ± 28.9	223.1 ± 32.3	256.8 ± 23.4
vehicle	72.2 ± 10.0	72.3 ± 9.2	72.6 ± 9.2	75.7 ± 18.9	73.9 ± 6.8	77.1 ± 11.8	75.8 ± 12.3	72.7 ± 12.3
voting	13.6 ± 3.6	4.8 ± 0.9	5.6 ± 1.4	2.0 ± 0.0	4.8 ± 1.0	9.0 ± 5.3	6.1 ± 2.8	8.9 ± 3.3
wine	6.1 ± 1.5	5.5 ± 1.1	5.5 ± 1.1	5.8 ± 0.6	6.1 ± 1.5	5.2 ± 0.4	5.9 ± 1.3	6.3 ± 0.7
zoo	8.4 ± 0.8	7.2 ± 0.9	7.1 ± 0.9	7.7 ± 0.5	7.7 ± 0.7	7.9 ± 0.9	7.0 ± 0.0	8.0 ± 0.5
overall	453.2 ± 1540.9	84.9 ± 229.4	101.8 ± 245.4	124.2 ± 343.6	102.1 ± 236.6	422.3 ± 1434.7	335.6 ± 1420	393.1 ± 1233.7

Appendix 2-9 Outputs of Two-way ANOVA for all assessments.

C4.5-p for diff CA: method vs data

Source	DF	SS	MS	F	P
method	6	2023.5	337.2	16.42	0.0000
data	32	7337.9	229.3	11.17	0.0000
Interaction	192	20845.2	108.6	5.29	0.0000
Error	2079	42694.1	20.5		
Total	2309	72900.7			
S=	4.5	R-Sq=	0.4144	R-Sq(adj)=	34.96%

C4.5-u for diff. CA: method vs data

Source	DF	SS	MS	F	P
method	6	2778.1	463.0	17.31	0.0000
data	32	7500.9	234.4	8.76	0.0000
Interaction	192	21227.2	110.6	4.13	0.0000
Error	2079	55621.0	26.8		
Total	2309	87127.2			
S=	5.2	R-Sq=	0.3616	R-Sq(adj)=	29.10%

C4.5-p for RR attribute: method vs data

Source	DF	SS	MS	F	P
method	6	127521.0	21253.5	196.83	0.0000
data	32	786048.0	24564.0	227.49	0.0000
Interaction	192	310906.0	1619.3	15	0.0000
Error	2079	224484.0	108.0		
Total	2309	1448959.0			
S=	10.4	R-Sq=	0.8451	R-Sq(adj)=	82.79%

C4.5-u for RR attributes: method vs data

Source	DF	SS	MS	F	P
method	6	102232.0	17038.7	203.9	0.0000
data	32	743290.0	23227.8	277.97	0.0000
Interaction	192	345245.0	1798.1	21.52	0.0000
Error	2079	173725.0	83.6		
Total	2309	1364492.0			
S=	9.1	R-Sq=	0.8727	R-Sq(adj)=	85.86%

C4.5-p for RR tree size: method vs data

Source	DF	SS	MS	F	P
method	6	4708053.0	784675.0	41.27	0.0000
data	32	46185564.0	1443299.0	75.92	0.0000
Interaction	192	155806654.0	811493.0	42.68	0.0000
Error	2079	39524422.0	19011.0		
Total	2309	246224692.0			
S=	137.9	R-Sq=	0.8395	R-Sq(adj)=	82.17%

C4.5-u for RR tree size: method vs data

Source	DF	SS	MS	F	P
method	6	427201.0	71200.1	30.49	0.0000
data	32	1403371.0	43855.4	18.78	0.0000
Interaction	192	2520081.0	13125.4	5.62	0.0000
Error	2079	4854389.0	2335.0		
Total	2309	9205042.0			
S=	48.3	R-Sq=	0.4726	R-Sq(adj)=	41.43%

NB for diff. CA: method vs data

Source	DF	SS	MS	F	P
method	6	4313.0	718.8	19.47	0.0000
data	32	27662.0	864.4	23.42	0.0000
Interaction	192	30959.0	161.2	4.37	0.0000
Error	2079	76742.0	36.9		
Total	2309	139675.0			
S=	6.1	R-Sq=	0.4506	R-Sq(adj)=	38.98%

NB for RR attribtues: method vs data

Source	DF	SS	MS	F	P
method	6	118223.0	19703.8	206.04	0.0000
data	32	902032.0	28188.5	294.76	0.0000
Interaction	192	343369.0	1788.4	18.7	0.0000
Error	2079	198820.0	95.6		
Total	2309	1562444.0			
S=	9.8	R-Sq=	0.8728	R-Sq(adj)=	85.87%

Appendix 2-10 Outputs of Tukey's test for Bon grouping.

AS methods		A. Difference of classification accuracy						B. Relative reduction of attribute						C. Relative reduction of decision tree size			
	subtract	C4.5-P		C4.5-U		NB		C4.5-P		C4.5-U		NB		C4.5-P		C4.5-U	
A	from A	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
CNS	CFS	-3.31	-0.76	-3.44	-0.66	-4.33	-0.81	-12.62	-1.62	-12.49	-1.74	-12.87	-1.37	-82.90	65.80	-39.28	-10.94
IG	CFS	-0.35	2.20	0.09	2.86	-0.65	2.87	-13.89	-2.90	-9.77	0.98	-13.07	-1.57	-77.00	71.70	-29.56	-1.22
RLF	CFS	-0.41	2.14	-0.23	2.55	-0.23	3.29	-15.56	-4.56	-10.53	0.22	-16.51	-5.02	-206.50	-57.80	-15.22	13.12
TNSP	CFS	-0.68	1.87	-0.37	2.41	-0.63	2.89	-0.29	10.70	1.54	12.29	-1.72	9.78	-75.40	73.30	-4.11	24.23
TNSU	CFS	-1.04	1.51	-0.66	2.12	-0.51	3.01	-3.00	7.99	0.84	11.59	-6.42	5.07	-70.60	78.10	-6.79	21.55
WRP	CFS	-1.34	1.20	-1.24	1.54	-2.48	1.04	6.17	17.17	6.30	17.05	5.92	17.42	-97.60	51.10	2.51	30.85
IG	CNS	1.69	4.24	2.13	4.91	1.92	5.44	-6.78	4.22	-2.65	8.10	-5.95	5.55	-68.40	80.30	-4.45	23.89
RLF	CNS	1.63	4.18	1.81	4.59	2.34	5.86	-8.44	2.56	-3.41	7.33	-9.40	2.10	-197.90	-49.30	9.89	38.23
TNSP	CNS	1.36	3.91	1.68	4.46	1.94	5.45	6.82	17.82	8.66	19.41	5.40	16.90	-66.80	81.90	21.00	49.34
TNSU	CNS	1.00	3.54	1.38	4.16	2.05	5.57	4.11	15.11	7.96	18.71	0.70	12.19	-62.00	86.70	18.32	46.67
WRP	CNS	0.69	3.24	0.81	3.59	0.09	3.61	13.29	24.29	13.42	24.16	13.04	24.54	-89.00	59.60	27.62	55.96
RLF	IG	-1.33	1.22	-1.71	1.07	-1.34	2.18	-7.16	3.83	-6.14	4.61	-9.20	2.30	-203.90	-55.20	0.17	28.51
TNSP	IG	-1.60	0.94	-1.84	0.94	-1.74	1.78	8.10	19.10	5.93	16.68	5.60	17.09	-72.70	75.90	11.28	39.62
TNSU	IG	-1.97	0.58	-2.14	0.64	-1.63	1.89	5.39	16.39	5.24	15.98	0.89	12.39	-68.00	80.70	8.60	36.95
WRP	IG	-2.27	0.28	-2.72	0.06	-3.59	-0.07	14.57	25.57	10.69	21.44	13.24	24.74	-95.00	53.70	17.90	46.24
TNSP	RLF	-1.55	1.00	-1.52	1.26	-2.16	1.36	9.77	20.76	6.70	17.45	9.05	20.54	56.80	205.50	-3.06	25.28
TNSU	RLF	-1.91	0.64	-1.82	0.96	-2.05	1.48	7.06	18.05	6.00	16.75	4.34	15.84	61.60	210.30	-5.74	22.61
WRP	RLF	-2.21	0.34	-2.40	0.38	-4.01	-0.49	16.23	27.23	11.46	22.20	16.69	28.19	34.60	183.20	3.56	31.90
TNSU	TNSP	-1.64	0.91	-1.69	1.09	-1.64	1.88	-8.21	2.79	-6.07	4.68	-10.45	1.05	-69.60	79.10	-16.85	11.50
WRP	TNSP	-1.94	0.61	-2.27	0.51	-3.61	-0.09	0.97	11.97	-0.62	10.13	1.89	13.39	-96.60	52.10	-7.55	20.79
WRP	TNSU	-1.58	0.97	-1.97	0.81	-3.72	-0.20	3.68	14.68	0.08	10.83	6.60	18.09	-101.40	47.30	-4.88	23.47

Appendix 2-11 Outputs of Kruskal-Wallis test.

Experiments		A. Difference of classification accurac				B. Relative reduction of attributes				C. Relative reduction of decision tree size			
Test classifier	Methods	Med	Mean rank	z value	z rank	Med	Mean rank	z value	z rank	Med	Mean rank	z value	z rank
C4.5-P	TNSP	0	1223.7	2.01	3	74.0	1314.7	4.68	2	9.5	1159.7	0.12	4
	TNSU	0	1168.1	0.37	4	70.6	1245.1	2.64	3	12.5	1219.4	1.88	2
	WRP	0	1163.4	0.23	5	78.3	1490.2	9.85	1	25.0	1378.8	6.57	1
	CFS	0	1068.7	-2.55	6	63.8	1148.9	-0.19	4	9.1	1160.0	0.13	3
	IG	0	1262.6	3.15	1	55.6	970.5	-5.44	6	2.1	1055.9	-2.93	6
	RLF	0	1236.6	2.39	2	55.0	922.8	-6.85	7	5.0	1080.4	-2.21	5
	CNS	0	965.2	-5.6	7	60.0	996.2	-4.69	5	2.5	1034.3	-3.57	7
	KW test	H=49.45, H=53.89 (adjusted for ties)				H=192.29, H=192.36 (adjusted for				H=62.48, H=63.47 (adjusted for ties)			
C4.5-U	TNSP	0	1226.5	2.09	3	73.3	1292.8	4.04	3	25.0	1324.4	4.97	2
	TNSU	0	1185.2	0.87	4	75.3	1297.8	4.19	2	16.7	1270.6	3.39	3
	WRP	0	1164.5	0.26	5	78.3	1441.9	8.42	1	41.6	1422.3	7.85	1
	CFS	0	1058.1	-2.86	6	63.8	1084.8	-2.08	4	12.5	1138.7	-0.49	4
	IG	0	1273.8	3.48	1	60.8	1028.9	-3.72	5	4.8	1044.3	-3.27	6
	RLF	0	1243.1	2.58	2	64.3	998.9	-4.61	6	12.5	1113.7	-1.23	5
	CNS	0	937.2	-6.42	7	60.0	943.4	-6.24	7	0.0	774.5	-11.21	7
	KW test	H=62.91, H=65.95 (adjusted for				H=156.99, H=157.03 (adjusted for ties)				H=202.18, H=203.28 (adjusted for ties)			
NB	TNSP	0	1233.0	2.28	4	73.3	1299.6	4.24	2				
	TNSU	0	1255.5	2.94	1	64.9	1172.3	0.49	3				
	WRP	0	1031.4	-3.65	6	78.3	1484.5	9.68	1				
	CFS	0	1083.9	-2.11	5	63.8	1157.3	0.05	4				
	IG	0	1244.6	2.62	3	52.9	1027.1	-3.78	5				
	RLF	0	1253.1	2.87	2	50.0	943.7	-6.23	7				
	CNS	0	986.9	-4.96	7	60.0	1003.9	-4.46	6				
	KW test	H=61.16, H=62.64 (adjusted for				H=158.45, H=158.5 (adjusted for ties)							

All tests have p-value of 0.000 for both hypotheses.

Appendix 2-12 Lower and upper 95% CI for means of difference of classification accuracy, relative reduction of attributes and relative reduction of decision tree size (in %) among AS methods, for all classifiers.

AS methods	A. Difference of classification accuracy						B. Relative reduction of attribute						C. Relative reduction of decision tree size			
	C4.5-P		C4.5-U		NB		C4.5-P		C4.5-U		NB		C4.5-P		C4.5-U	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
TNSP	-0.22	0.63	0.13	1.23	0.61	2.24	63.92	68.78	65.87	70.25	62.54	67.82	9.04	19.49	31.98	39.73
TNSU	-0.60	0.27	-0.13	0.90	0.74	2.34	61.03	66.26	64.94	69.79	57.76	63.19	15.68	22.41	29.25	37.11
WRP	-1.17	0.25	-0.93	0.55	-1.27	0.43	70.77	74.87	70.77	74.87	70.77	74.87	-34.08	18.14	38.26	46.69
CFS	-0.85	0.06	-0.90	0.22	-0.41	1.00	58.94	63.35	58.94	63.35	58.94	63.35	10.66	19.90	22.13	29.46
IG	0.09	0.97	0.58	1.69	0.83	1.99	49.87	55.63	53.86	59.65	50.72	56.94	9.30	16.03	-3.44	24.26
RLF	-0.01	0.95	0.27	1.36	1.06	2.60	48.18	54.00	53.10	58.88	47.24	53.52	-204.93	-28.82	21.06	28.43
CNS	-3.43	-1.43	-3.36	-1.41	-3.42	-1.11	51.17	56.89	51.17	56.89	51.17	56.89	0.43	13.03	-5.91	7.28

Appendix 2-13 Outputs of t-test for paired two sample for means* of pruned and unpruned.

		Difference of CA		RR of attributes		RR of decision tree size	
Test classifier	AS method	Mean	SD	Mean	SD	Mean	SD
C4.5-P	none	82.45	13.09	31.82	47.34	88.0	222.0
C4.5-U	none	81.84	13.40	31.82	47.34	453.0	1541.0
	Difference	0.61	1.55	0.00	0.00	-365.0	1436.0
C4.5-P	TNSP	82.65	12.74	7.29	5.07	75.2	209.5
C4.5-U	TNSP	82.52	12.70	7.21	5.36	84.9	229.4
	Difference	0.13	1.58	0.08	1.76	-9.7	57.1
C4.5-P	TNSU	82.29	13.13	7.90	5.64	73.6	207.7
C4.5-U	TNSU	82.22	13.12	7.56	6.38	101.8	245.4
	Difference	0.07	1.53	0.33	3.01	-28.2	96.0
C4.5-P	WRP	81.99	12.58	5.57	3.24	65.7	202.4
C4.5-U	WRP	81.64	12.92	5.57	3.24	124.2	343.6
	Difference	0.34	0.98	0.00	0.00	-58.4	234.7
C4.5-P	CFS	82.06	12.35	8.52	5.74	66.0	202.6
C4.5-U	CFS	81.50	12.58	8.52	5.74	102.1	236.6
	Difference	0.56	1.29	0.00	0.00	-36.1	75.6
C4.5-P	IG	82.98	12.44	11.95	11.14	78.0	213.0
C4.5-U	IG	82.97	12.61	11.68	13.09	422.0	1435.0
	Difference	0.01	1.81	0.27	2.86	-344.0	1381.0
C4.5-P	RLF	82.92	12.41	11.50	7.61	82.0	213.0
C4.5-U	RLF	82.65	12.51	10.75	8.96	336.0	1420.0
	Difference	0.27	1.43	0.76	3.20	-254.0	1329.0
C4.5-P	CNS	80.02	13.74	9.02	4.60	73.0	208.0
C4.5-U	CNS	79.45	13.85	9.02	4.60	393.0	1234.0
	Difference	0.57	2.14	0.00	0.00	-320.0	1175.0

Difference = Pruned-Unpruned (n =33)

*Input values are a mean value of each data based on 10-fold cross validation.

Appendix 2-14 Processing time (in seconds) for all attribute selection methods.

Each letter, P, U or N, inside the bracket indicates the method of testing algorithms, pruned and unpruned, or naïve Bayes, respectively.

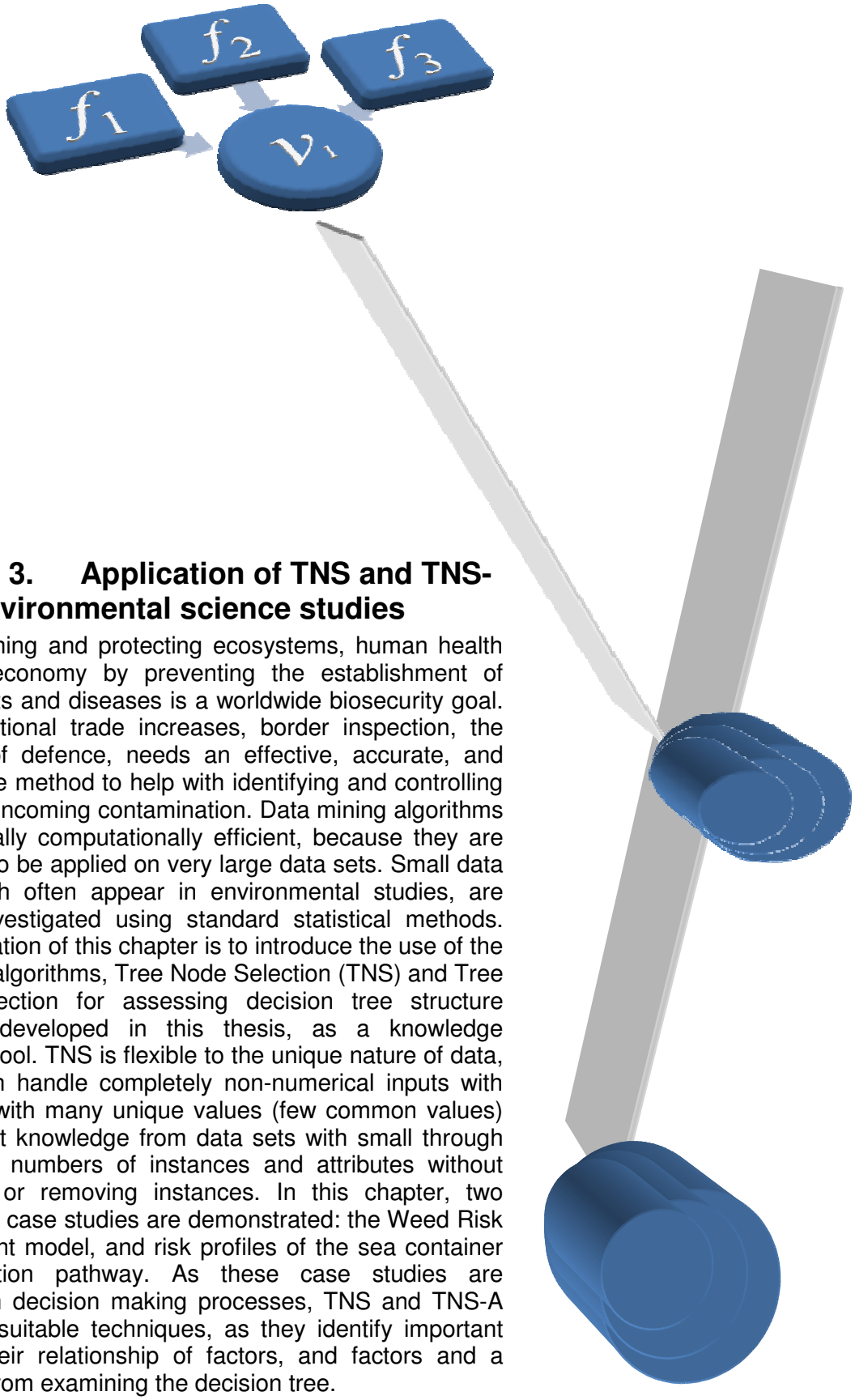
Data	(P) TNSP	(U) TNSP	(N) TNSP	(P) TNSU	(U) TNSU	(N) TNSU	CFS	CNS	(P) IG	(U) IG	(N) IG	(P) RLF	(U) RLF	(N) RLF	WRP
adult	2375.45	2048.67	212.28	2588.54	2119.77	212.77	3.99	64.61	1919.31	1682.86	172.81	5422.00	5308.38	3842.41	5313.16
anneal	8.03	6.73	5.30	12.60	9.69	6.47	0.39	0.62	56.47	44.17	24.12	57.85	45.83	26.93	233.43
arrhythmia	64.00	57.54	36.11	79.40	71.55	43.86	1.36	7.99	2630.29	2549.90	1280.61	2856.95	2778.75	1236.42	1279.50
audiology	7.09	6.58	5.91	10.94	9.93	8.51	0.33	0.62	44.79	41.59	34.17	44.80	41.98	34.66	60.46
balance-scale	2.65	2.48	2.11	2.63	2.52	2.17	0.28	0.32	2.53	2.33	1.93	2.80	2.63	2.27	2.14
bands	0.98	0.97	0.90	1.68	1.64	1.38	0.38	0.97	36.30	37.23	23.55	35.15	35.88	24.49	116.29
breast-cancer2	1.05	1.04	1.04	4.19	4.18	3.60	0.30	0.32	4.14	4.10	3.52	4.26	4.25	3.59	1.37
credit-screening	4.88	4.72	3.94	8.60	8.20	6.22	0.34	0.45	9.99	9.60	7.17	10.30	9.80	7.57	1.58
ecoli	3.60	3.44	3.11	3.62	3.38	3.08	0.28	0.32	4.00	3.69	3.39	4.23	3.94	3.55	3.03
flags	8.65	7.98	7.02	11.27	10.26	9.28	0.32	0.37	15.87	14.69	13.03	15.76	14.60	13.48	18.38
german credit	14.16	12.92	8.32	18.68	17.88	10.99	0.35	0.77	18.60	17.66	10.79	20.16	19.41	12.79	40.85
heart-c	6.54	6.14	5.23	6.44	6.10	5.25	0.30	0.34	6.83	6.49	5.55	6.77	6.65	5.60	1.48
hepatitis2	1.88	1.83	1.73	5.24	5.07	4.46	0.30	0.31	8.48	8.12	7.31	8.19	8.09	6.77	2.13
horse-coli	3.19	3.06	2.81	9.09	8.77	7.10	0.28	0.37	13.10	12.99	9.87	13.47	13.07	10.10	12.17
hypothyroid	11.09	10.53	7.87	11.05	10.47	7.90	0.68	1.57	60.88	58.25	36.94	106.57	103.75	76.45	181.88
ionophere	6.72	6.57	5.23	6.70	6.40	5.20	0.38	0.47	29.99	28.62	17.78	30.80	29.41	18.66	14.36
iris	1.07	1.00	1.00	1.09	1.04	1.00	0.26	0.26	1.78	1.67	1.66	1.85	1.75	1.71	0.77
kr-vs-kp	33.34	26.65	17.88	49.44	39.34	26.15	0.65	1.44	67.89	55.48	34.87	100.91	87.33	67.78	50.95
labor-relations	0.83	0.84	0.82	1.41	1.37	1.31	0.25	0.25	5.46	5.37	5.07	5.51	5.37	5.12	0.86
letter recognition	1774.77	1634.06	358.61	1753.17	1612.34	359.27	2.62	33.06	1730.67	1610.09	350.30	2545.79	2415.73	1167.48	9391.35
mammographic	1.99	1.96	1.75	3.45	3.34	2.74	0.34	0.33	3.26	3.13	2.61	4.06	3.91	3.34	2.29
mushroom	7.74	6.59	5.13	7.69	6.62	5.08	0.69	1.78	47.99	41.59	29.37	189.16	182.12	170.12	73.87
pimadiabetes	5.72	5.28	4.22	5.67	5.30	4.15	0.34	0.37	5.37	5.15	4.01	6.06	6.11	4.67	7.13
primary-tumor	8.38	7.61	5.98	8.42	7.65	5.94	0.28	0.39	10.24	9.44	7.42	10.64	9.71	7.39	4.97
segment	13.24	11.97	9.00	13.23	11.87	9.04	0.38	0.57	22.79	20.67	14.77	24.29	21.89	16.46	68.72
sonar	5.77	5.57	4.90	5.82	5.62	4.82	0.37	0.57	47.93	46.56	31.45	48.64	47.10	31.36	10.47
soybean	19.61	16.78	11.72	24.67	21.13	14.14	0.38	0.94	32.21	28.24	18.58	34.06	29.98	20.89	153.09
splice	17.59	15.12	8.80	78.71	69.53	37.35	1.31	12.00	307.99	295.48	134.47	197.45	182.04	122.52	360.63
vehicle	22.00	18.73	10.85	22.14	18.66	10.86	0.40	0.62	24.16	20.48	11.33	26.34	22.36	12.97	60.11
voting	2.41	2.43	2.21	3.73	3.65	3.32	0.29	0.37	7.47	7.30	6.31	7.75	7.57	6.58	0.97
waveform	211.88	189.30	48.41	211.69	188.48	48.22	1.21	8.00	210.12	188.19	46.53	276.23	253.82	112.35	611.01
wine	2.86	2.81	2.59	2.88	2.79	2.58	0.28	0.32	6.07	5.98	5.42	6.12	6.09	5.42	3.63
zoo	3.29	3.21	3.20	3.24	3.09	3.07	0.30	0.32	6.54	6.40	6.28	6.48	6.21	6.32	2.13

Appendix 2-15 Outputs of One-way ANOVA test (top) and description statistics (bottom) for the subset evaluator approaches.

```
Source  DF       SS        MS     F      P
Factor   2    6548476   3274238  2.89  0.060
Error   96  108808356   1133420
Total   98  115356832

S = 1065   R-Sq = 5.68%   R-Sq(adj) = 3.71%


                        Individual 95% CIs For Mean Based on Pooled StDev
Level   N  Mean  StDev     +---------+---------+---------+---------
CFS    33     1      1     (---------*----------)
CNS    33     4     12     (---------*----------)
WRP    33   548   1844                   (----------*---------)
                          +---------+---------+---------+---------
                        -350         0       350       700

Pooled StDev = 1065
```

Appendix 2-16 Attribute selection processing time (in seconds) for the ranking filter approach excluding evaluating the test algorithms.

Data	TNSP	TNSU	IG	RLF
adult	28.72	26.26	3.22	3694.58
anneal	0.63	0.58	0.37	3.27
arrhythmia	3.26	3.17	0.83	8.88
audiology	0.57	0.57	0.37	0.83
balance-scale	0.45	0.46	0.28	0.58
bands	0.52	0.51	0.40	1.67
breast-cancer2	0.36	0.36	0.30	0.37
credit-screening	0.53	0.51	0.32	1.13
ecoli	0.46	0.42	0.27	0.48
flags	0.47	0.47	0.32	0.47
german credit	0.63	0.61	0.37	2.52
heart-c	0.48	0.43	0.27	0.48
hepatitis2	0.42	0.42	0.27	0.37
horse-coli	0.49	0.49	0.29	0.58
hypothyroid	0.83	0.79	0.56	40.94
ionophere	0.56	0.52	0.37	0.87
iris	0.32	0.32	0.25	0.32
kr-vs-kp	0.73	0.70	0.56	33.39
labor-relations	0.30	0.30	0.25	0.26
letter recognition	15.98	14.74	1.98	824.51
mammographic	0.43	0.44	0.32	1.02
mushroom	0.77	0.73	0.62	141.44
pimadiabetes	0.50	0.45	0.32	0.97
primary-tumor	0.49	0.46	0.32	0.57
segment	0.58	0.55	0.43	2.08
sonar	0.51	0.50	0.37	0.67
soybean	0.57	0.57	0.38	2.28
splice	1.22	1.18	0.69	55.06
vehicle	0.69	0.66	0.37	2.02
voting	0.40	0.35	0.32	0.57
waveform	2.65	2.45	0.89	66.69
wine	0.40	0.40	0.27	0.37
zoo	0.42	0.37	0.27	0.32



Chapter 3. Application of TNS and TNS-A for environmental science studies

Maintaining and protecting ecosystems, human health and the economy by preventing the establishment of exotic pests and diseases is a worldwide biosecurity goal. As international trade increases, border inspection, the first line of defence, needs an effective, accurate, and quantitative method to help with identifying and controlling unwanted incoming contamination. Data mining algorithms are generally computationally efficient, because they are designed to be applied on very large data sets. Small data sets, which often appear in environmental studies, are usually investigated using standard statistical methods. The motivation of this chapter is to introduce the use of the computer algorithms, Tree Node Selection (TNS) and Tree Node Selection for assessing decision tree structure (TNS-A), developed in this thesis, as a knowledge discovery tool. TNS is flexible to the unique nature of data, e.g., it can handle completely non-numerical inputs with attributes with many unique values (few common values) and extract knowledge from data sets with small through very large numbers of instances and attributes without modifying or removing instances. In this chapter, two biosecurity case studies are demonstrated: the Weed Risk Assessment model, and risk profiles of the sea container contamination pathway. As these case studies are involved in decision making processes, TNS and TNS-A would be suitable techniques, as they identify important factors, their relationship of factors, and factors and a decision, from examining the decision tree.

Study I. Application of Tree Node Selection and Ant-Miner algorithm for the weed risk assessment model (Fukuda and Brown 2007a,b)

3.1. Introduction

This chapter covers the brief concept of the weed risk assessment (WRA) model and its background, and introduces how the application of the attribute selection method, Tree Node Selection (TNS), developed in Chapter 2, can help the process for the WRA model by identifying important or key questions to improve the efficiency of the assessment process for alien plants as a knowledge discovery tool. In comparison to TNS, an interesting data mining approach, Ant-Miner, is introduced. Also discussed are results from TNS and Ant-Miner analyses, which provided knowledge on how different alien plants are assessed for risk differently among different countries and climates by comparing the Australia and New Zealand WRA model and the Hawaii and Pacific WRA model. It is important to note that the results were obtained by applying a knowledge discovery tool on the case studies, with the aim of understanding the structure of the data by extracting information, rather than prediction or forecasting. This study aimed to inspire various environmental scientists about newly collected data, i.e., the data were not large or sophisticated enough to carry out the prediction model for assessing the risk of new plants or contaminated containers (a future goal) by extracting or adding insights to their data to help with the future decision making process (as previously mentioned in Chapter 1).

3.1.1. The Weed Risk Assessment (WRA) model

Effective strategies to mitigate, control existing and future invasive organisms are important for maintaining and protecting our healthy ecosystem. The entrance and spread of invasive weeds (alien plants) can threaten the native environment. They can alter the fundamental structure of the ecosystem by changing its composition, structure, and function (Yeates and Williams 2001). The weed risk assessment (WRA) model (Pheloung et al. 1999) provides an informed decision prior to introducing potentially invasive plant species into the country. The WRA was established as a biosecurity tool to evaluate new plants prior to introduction in Australia, and has been tested and modified to adapt to the unique climate and environment of different countries, for example, New Zealand (Pheloung et al. 1999) and Hawaii and the Pacific Islands (Daehler et al. 2004), referred to here as Hawaii/Pacific.

The WRA models have 49 questions on the impacts of weeds to allow assessment of their weediness (see the blank WRA sheet in Appendix 3-1; Pheloung et al. 1999). Individual plant species are assessed by answering questions in the WRA model, resulting in a score from -14

(benign taxa) to 29 (maximum weediness). The total score is then evaluated into three possible recommendations: accept the plant for import (score < 1), further evaluation required for the plant (score from 1-6), and reject the plant for import (score > 6). Additionally, in the WRA for Hawaii/Pacific, a second screening process is applied for scores from 1-6 to determine a further recommendation to either accept or reject (see detailed criteria in Daehler et al. 2004). Daehler et al. (2004) found from a comparison between the WRA and experts' opinions, the second screening process for the WRA improves the number of correctly identified non-pests, i.e., non-pest classification accuracy with the second screening is improved to 85% from 66% without, as well as classifying additional minor pests as non-pests.

The WRA model is beneficial as a decision making tool, since it eases the border security process of plant risk assessment. However, some key issues are of concern to set up such a model. For example, the WRA process is not part of any legal process to prevent importing, except in the USA if the plant is stated in the State or Federal Noxious Weed List (Daehler et al. 2004). Minimising biases is important as personal opinion on assessing *invasiveness* of weeds can vary among different fields of expertise (Pheloung et al. 1999). Besides the WRA models, numerous different approaches have been taken to predict invasion or potential distribution of species, which were generally investigated by statistical methods, e.g., multiple regression, regression trees, discriminant analysis or other multivariate techniques, or biologically, e.g., habitat modelling (Rejmanek 2001), although Williamson (2001) commented that "*Prediction is not the same as explanation, nor the same as understanding. Many of the attempts at predicting invasive impacts are, to my mind, much more explanations of impacts…The impact of invasions cannot, in general, be predicted but can be subject to risk assessment, leading to policies for risk avoidance and risk containment."* Hence, this study proposes a knowledge discovery concept, data mining techniques, for the weed risk assessment process, aiming to produce knowledge about data via the use of a tool that describes the phenomena. This concept is not a prediction tool; it is to achieve understanding and increasing knowledge about the WRA model itself.

3.1.2. Application of TNS to identify important WRA questions

In this study, the attribute selection method developed earlier in this thesis, Tree Node Selection (TNS), described in Chapter 2, was used to select subsets of questions (attributes) for the Australia and Hawaii/Pacific WRA models as a knowledge discovery tool, as it was found to have the most consistent performance to select good predictors or attributes, sustaining the classification accuracy among all other existing attribute selection methods.

The purpose of this study and applying the TNS method is to help future WRA by identifying important and unimportant questions in the decision making process in the alien plant importation and biosecurity border security system. TNS is particularly suitable here, as it detects important factors by directly investigating the decision making system for the weed risk assessment process; the input source for TNS is a pre-generated decision tree. Note that it could be suspected that results of TNS could be biased to the decision tree structure, but it was found that TNS fared well in comparison to other attribute selectors in selecting attribute subsets on standard test data, and the attributes selected by TNS were valid even for different learning schemes such as naïve Bayes (see details in Chapter 2).

The usefulness of TNS results in this study will be, if a particular question is found to be more important than other questions for judging high-risk plants, then this question may be highlighted as important to answer, or preferably answer accurately by collecting as many resources as possible. Note that *attributes* in this context are actually questions in the WRA model. If it is impossible to answer the questions because the species is new or there is a lack of resources for the new environment, then the questions identified by such an attribute selection method may be divided into a few specific detailed questions, changing their aspects to make them easier to answer. On the other hand, questions that are found to be less important may be able to be removed from the WRA system. At the same time, if the question is too difficult to answer, then the plant is classified as *evaluate* or *more information required* (as answers tend to remain blank). If some particular questions are more likely to be unanswered and if the investigation from this study can identify which questions, then it would be best to narrow or even remove the types of question in order to increase the accuracy of identifying the risk of the plant. Note that the study data sets from Australia and Hawaii/Pacific contained less than 20% and 10% respectively of *evaluate* or *more information required* responses. Thus, removing a few questions that appear to classify as *evaluate* or *more information required* decisions may not impact on the overall assessment. Hence, it is important to understand the model as it may further increase classification accuracy and improve the WRA process; this can be achieved or improved by uniquely applying the attribute selection method.

3.1.3. Ant-Miner as the attribute selection approach

In addition to TNS, this study introduced the unique data mining tool, Ant-Miner (Parpinelli et al. 2002) as an attribute section tool, developed based on Ant Colony Optimisation (ACO). ACO is a metaheuristic inspired by the foraging behaviour of ant colonies, i.e., tracking of pheromones, with the objective of solving discrete optimisation

problems, developed in the 1980s by Dorigo and Stützle (2004). Generally, due to its nature, ACO has been applied to the travelling salesman problem, and various other fields (sequential ordering, flow shop scheduling and the graph colouring problem; details in Dorigo and Stüzle 2004), though its application in environmental science is still uncommon. An interesting comparison between TNS and Ant-Miner is that their heuristic functions are different; TNS based on C4.5 decision tree algorithms computes entropy for attributes as a whole whereas Ant-Miner computes it for attribute-value pairs only (Parpinelli et al. 2002). Practically, ACO identifies the shortest pathway to classify each plant species (as either *accept*, *evaluate* or *reject*), whereas TNS selects the most important or predictive attributes using a pre-generated decision tree as the information source.

This chapter describes an application of TNS to WRA, briefly introduces the Ant-Miner algorithm, and finally presents the ACO algorithm. Another interesting comparison between TNS and Ant-Miner is that TNS identifies the most important questions that appear from constructing the decision tree (since TNS uses the pre-generated C4.5 decision tree as its information source), while Ant-Miner detects the shortest pathway to predict plant risk by selecting the fewest nodes (WRA questions). Results describe how the different WRA systems and questions were considered to be important between Australia and Hawaii/Pacific and how the selection process of questions is different between TNS and Ant-Miner. This study provides knowledge by unique data mining approaches to help plan the cost and time effective WRA model for the future.

3.2. Data and methods

3.2.1. Data set

The data set is taken from the website of the Institute of Pacific Islands Forestry, Pacific Island Ecosystems at Risk (PIER 2007); http://www.hear.org. The data source shows two types of risk assessments on WRA models: risk assessments for species that are listed on PIER, and not listed on PIER. Both sets of data have the score for a single plant species that is assessed by the Australia and Hawaii/Pacific WRA models; 163 and 555 plants are assessed by the Australia and Hawaii/Pacific models respectively. The original WRA

Table 3-1 Proportion of plant classes for Australia (*n*=163) and Hawaii/Pacific (*n*=555) WRA model.

Class	**Reject**	**Accept**	**Evaluate/More information**
Australia	131 (80%)	3 (1%)	20 (13%) for evaluate 9 (6%) for more information
Class	**High risk**	**Low risk**	**Evaluate**
Hawaii and Pacific	176 (32%)	321 (58%)	58 (10%)

(Australian) questionnaire blank sheet, developed by Pheloung et al. (1999) is shown in Appendix 3-1, and the data source and Hawaii/Pacific WRA are accessible from Daehler et al. (2004). Both WRA models have 8 sections, which are divided into several questions, with a total of 49 questions. Some questions, e.g., 4.10 from WRA, differ between the two models, as the Hawaii/Pacific model was adjusted from the Australian model.

As previously discussed in the introduction section, the total score for each plant is categorised as a class. The Australia model has four classes: *reject* (score > 6), *evaluate* (1 to 6), *more information required* (score > 4, but majority of questions unanswered) and *accept* (< 1). The Hawaii/Pacific model has three classes: *high risk* (> 6), *evaluate* (1-6) and *low risk* (< 1). Note that the Hawaii/Pacific model has a second screening process for the class *evaluate*, but the second screening process is not used in this study; such plants are left classified as *evaluate*. Table 3-1 shows the proportion of each class. The Australian model classified most of the plant species as *reject* (80%) and very few as *accept* (3%), since these data were collected to focus on high risk plants (Williams, personal communication, 17 Sep, 2007), whereas the class distribution is relatively balanced for Hawaii/Pacific: 32% for *high risk*, 58% for *low risk* and 10% for *evaluate*.

3.2.2. Tree Node Selection method

The detailed algorithm for the Tree Node Selection method (TNS) was described in Chapter 2 (Section 2.2.3.1), thus is not shown in this chapter. For the data processing, firstly the input data were separated into training (60%) and test data (40%), and the training data were processed by the C4.5 decision tree algorithm (Quinlan 1993) to create the input decision tree (information source) for TNS. An evaluation process selected the best subset of attributes (questions), which was tested on the test data, to obtain the classification accuracy for the selected subset of attributes. As this study focused on understanding the data structure (WRA system) rather than creating the prediction or classification rule, 60% of the data (training set) were used to represent the data structure for extracting key questions.

From Section 2.2.3.1 in Chapter 2 and equation 2-3: $I(a_i) = \Sigma\, I\,(v_k)\ \forall\ v_k \mid L(v_k) = a_i \in A$, in Chapter 2, each attribute a_i is ranked with the overall total instances, calculated from the sum of total instances at a node v_i (labelled as $L(v_k) = a_i$, see details in Section 2.2.3.1).

In this study, TNS results were reported from determining the *proportion of total instances* (in %), *Prop I* (a_i). While $A = \{a_1, \ldots, a_{na}\}$ and na ($1 \leq i \leq na$) is the number of unique attributes (see Section 2.2.3.1), the *proportion of total instances* detected by TNS for a_1 is defined by

$$Prop\ I\ (a_1) = \frac{I(a_1)}{\sum_{i=1}^{na} I(a_i)}. \quad (3\text{-}1)$$

Additionally, note that the approach that is taken for the following section, Chapter 3, Study II: Sea container risk profile investigation, uses three-fold cross-validation to achieve more detailed investigations for the decision tree structures with TNS and TNS for assessing decision tree structure (TNS-A).

3.2.3. Ant-Miner program

Generally, Ant-Miner produces *N* solutions or paths with an overall classification accuracy for *N*-fold cross validation. In this study, Ant-Miner classification accuracy was obtained from *N*-fold cross validation (*N*=10). To allow an interpretation of the path, a single solution (path) was obtained for the whole data set. Thus, firstly, the Australia and Hawaii/Pacific WRA model was analysed separately using 10-fold cross validation with four different parameter settings; three parameters (*min cases per rule* = 10, *max uncovered cases* = 10, and *no rules converg* = 10) were kept the same, but the number of ants was changed to 50 and 100, and applied to iterations of 25 and 100, respectively. The parameter settings that provided the best-represented results, i.e., the highest classification accuracy, were used to obtain the classification accuracy.

Secondly, the Ant-Miner program, developed by Parpinelli et al. (2002) detects classification rules using only 10-fold cross validation. However, in this study, the program was modified to provide a single classification rule, based on the entire data set. This classification rule was then used to understand the structure of the shortest pathway.

3.2.4. Ant Colony Optimization

The Ant Colony Optimization (ACO) algorithm is swarm intelligence that is generated by mimicking real ant behaviour. Ants write, read and estimate the amount of pheromone trail (proportional to the utility of using a particular arc) to build a good solution (Dorigo and Stützle 2004). The stronger the pheromone trail, the higher its desirability. Ants follow a probabilistic decision biased by the amount of pheromone. If no pheromone trail exists, ants move randomly (García-Martínez and Herrera 2007). A brief explanation of the Simple ACO (S-ACO) algorithm follows.

Let $G = (N, A)$ be the graph to each arc (i, j), and an associated variable τ_{ij}, the pheromone trail. Assume all the arcs A have a constant amount of pheromone (τ_{ij} =1, $\forall(i, j) \in A$) at first. Then, a probability P is defined for an ant k travelling from a node i to the next node j using τ_{ij} as follows,

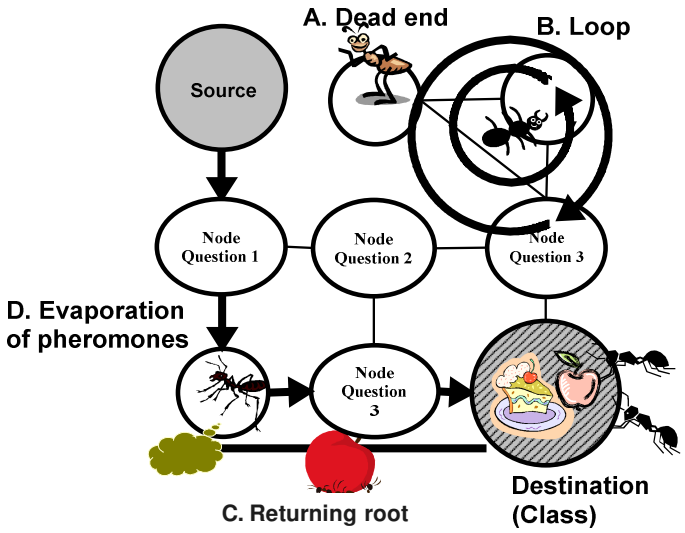


Fig. 3-1 Diagram of ants building a solution.

$$P_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}}{\sum_{l \in N_i^k} \tau_{ij}^{\alpha}}, & \text{if } j \in N_i^k \\ 0, & \text{if } j \notin N_i^k \end{cases}, \tag{3-2}$$

where α ($\in \{s, l\}$ when s and l are short and long branches respectively) is a parameter defining the relative importance weight of the pheromone trail, and N_l^k is the neighbourhood of ant k in node i that contains all the nodes directly connected to node i in the graph $G = (N, A)$, but excludes the predecessor of node i (the last node that the ant visited before moving to i) so as to avoid the ants returning to the node they visited immediately before node i. When N_l^k is empty (a dead end, as shown for example in Fig. 3-1, A), node i's predecessor is included into N_l^k. During this process, ants receive pheromone several times by going back and forth; consequently, this can lead to loops (seen in Fig. 3-1, B). Loop elimination is carried out by an iterative scanning process; the path from the destination node back to a given node is scanned. If another instance of the node is reached along the way, the subpath from this instance back to the original instance of the node is a loop, which can be eliminated.

Let a change of amount of pheromone be $\Delta\tau^k$, deposited by the k^{th} ant on arc (i, j) that is visited during their return travel (Fig. 3-1, C),

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta\tau^{k}. \tag{3-3}$$

When an ant deposits pheromone earlier than one travelling a longer path, it deposits more pheromone on the shorter path. At the same time as updating the pheromone trail, pheromone trail evaporation (Fig. 3-1, D) is considered, to avoid all ants moving toward a suboptimal path by converging; losing pheromone intensity favours the exploration of different paths. Let ρ be a parameter, where $\rho \in (0, 1]$, then when ant k moves between nodes, the pheromone trails are evaporated as

$$\tau_{ij} \leftarrow (1 - \rho)\tau_{ij}, \qquad \forall (i, j) \in A. \tag{3-4}$$

A complete cycle of an iteration of ACO involves pheromone evaporation and deposition, and ant movement.

3.2.5. Ant-Miner algorithm

The following section briefly introduces the main theoretical modifications of Ant-Miner from the ACO algorithm. Ant-Miner is similar to the decision tree algorithm, such as C4.5 (Quinlan 1993) that discovers classification rules by following a divide-and-conquer approach:

IF < *term*1 and *term*2 and ...> *THEN* <*class*>

As previously mentioned, the heuristic functions for decision tree algorithms and Ant-Miner differ in how they consider the entropy; for the former they are computed for an attribute as a whole, but the latter computes them for an attribute-value pair only (Parpinelli et al. 2002).

The procedure of discovering classification rules is following two conditions. Firstly, an ant starts with an empty rule and adds one term at a time to its current partial rule until one of the two following conditions is satisfied:

1) Adding any term to the rule would result in it covering less than a user-specified minimum number of cases.
2) All attributes have already been used by the ant to create the rule antecedent.

Secondly, the rule can be pruned to eliminate irrelevant terms and thirdly, the amount of the pheromone is increased in the trail followed by the ant and decreased elsewhere (evaporation). Then, newly updated pheromone guides other ants to construct the rule until one of the following is satisfied:

1) The number of constructed rules is equal to or greater than the user-specified number of ants.
2) When the exact same rule has been created by a user-specified number of successive ants.

The algorithm is described in detail in Parpinelli et al. (2002). To operate a data mining algorithm, Ant-Miner modifies the P_{ij} function (originally equation 3-2 from ACO) which allows the current ant to iteratively add one term at a time to its current partial rule. Let η_{ij} be a value of the heuristic function to estimate the quality or precise value of the entropy associated with the arc (i, j) to improve the predictive accuracy of the rule in equation 3-5, where I is the total number of attributes, J_i is the number of values in the domain of the i^{th}

attributes and x_i is set to 1 if the attribute Ai was not yet used by the current ant, or to 0, otherwise.

$$P_{ij}^{k} = \frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i=1}^{I} x_i \sum_{j=1}^{J_i} (\eta_{ij}\tau_{ij}(t))} \tag{3-5}$$

Pheromone updating (equation 3-3 for the ACO) is calculated from:

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^{I} Ji}, \tag{3-6}$$

which is inversely proportional to the number of values of all attributes. Then, the pheromone update can be carried out by increasing and decreasing for arcs that are used or not used, respectively (details in Parpinelli et al. (2002) and the equation 3-4 from the ACO). Ant-Miner parameters are defined by the experiments of running a few different parameter settings, and the best results, e.g., higher classification accuracy, are introduced in the following sections. Note that all classification rules are pruned.

3.3. Results and discussion

3.3.1. Classification accuracy

Table 3-2 and Table 3-3 show the classification accuracy obtained respectively from TNS (shown as confusion matrices) and Ant-Miner via the best parameter setting (the number of ants and iterations are 100). The classification accuracy (CA) of TNS (with best subset of selected questions) shows that Hawaii/Pacific (84%) has higher CA than Australia (79%). Similarly, Ant-Miner obtained reasonably higher classification accuracy for Hawaii/Pacific (80.15 ± 1.24%) than Australia (71.02 ± 2.26%). The higher classification accuracy for Hawaii/Pacific may be due to its more balanced class distribution compared to the Australia WRA data, where the high proportion of instances with class *reject* may have made prediction more difficult. However, obtaining over 71% classification accuracy in a real case study by C4.5 with TNS, and even the experimental Ant-Miner algorithm, can be said to be reasonable. Hence, it could be said that the structure of the data is reasonable enough to bring representative and reliable results in this investigation. However, as the CA for the Australia WRA data (CA=79% for TNS) is similar to the original proportion of the *reject* class (CA=80% in Table 3-1), it is possible that the decision tree was biased towards *reject*; the distribution of *reject* instances in the confusion matrix in Table 3-3 shows that most of the instances (thus most of the errors) were classified as *reject*. In fact, the data set contained mainly high risk plants, as previously discussed. This study was conducted to understand the decision pathways rather than to produce a classification decision tree for use as a prediction

tool. Thus, results for the Australia WRA represent decision pathways mainly for *reject*ed plants (and perhaps also plants classified *evaluate*, since there were 5 correctly classified *evaluate* instances).

3.3.2. Australia WRA system

The relative usage of key questions selected by TNS, indicated by the WRA question section number (details in Appendix 3-1) and a brief summary, are shown in Table 3-2. Each percentage in Table 3-2 indicates the TNS decision proportion (in %), *Prop I* (a_i) value, described in Section 3.2.2. For example, when the final class of the investigated plant species was *reject* or *evaluation required*, the question about the buoyancy of the propagules (section 7.05 in the WRA model in Appendix 3-1) accounts for respectively 73% and 24% of the decisions in the decision tree (*Prop I* (a_i) value), thus is very important for the classification of such species. On the other hand, when the final class is *accepted*, then 100% of the time, the plant's climate and distribution process, shown as Section 2.03: *broad climate suitability* (*environmental versatility*) is assessed to make this decision.

For the results of Ant-Miner, Fig. 3-2 shows the shortest pathway and nodes for the selected key questions. For example, three common questions were detected among classes; if reproduction by *self-fertilisation* (6.04) is unknown, this connects to the classes *reject* and *evaluation*, if the *vegetative propagation reproduction* (6.06) is true (*yes*) or unknown, this connects to *reject* and *evaluation* respectively, and if the minimum generative time for reproduction is one year (6.07) or unknown, this connects to *reject* and *more information required*, respectively. Besides the above, three pathways were detected for high risk plants (*reject*), when the *plant is beyond native* (3.01), there is no evidence of substantial reproductive failure in the *native habitat* (6.01), and the plant is unknown as *a host for recognized pests and pathogens* (4.06). The pathways for *more information required* were created by all questions – *weedy races* (1.03), *minimum generative time* (6.07), *wind dispersal* (7.04) and *herbicide control* (8.03) – which are all unanswered (indicated by a question mark [?] in Fig. 3-2); this is a reasonable finding, as more unanswered questions lead to requiring more information about the plant.

This may suggests that these questions may need to be improved by adding more specific questions to help answer them. If these questions are in fact difficult to answer, perhaps even removing them may help the overall analysis, though note that it is important to keep the question about the *minimum reproduction time* (6.07), which was found to be important for judging the class.

Table 3-2 Australia WRA key questions selected by TNS, with TNS decision proportions for each class shown as percentages (*Prop I* (a_i) value).

Australia	Section 2	Section 3		Section 4	Section 5	Section 7		
	Climate /Distribution	Weed elsewhere		Undesirable traits	Plant type	Dispersal mechanisms		
	2.03: Broad climate suitability	**3.01: Naturalised beyond native**	**3.03: Agr./hort./ forestry**	**4.04: Unpalatale to grazing animals**	**5.05: Nitrogen fixing woody plant**	**7.02: Dispersed intentionally by people**	**7.04: Adapted to wind dispersal**	**7.05: Buoyant**
Eval.	0%	0%	24%	24%	0%	0%	28%	24%
More	0%	0%	17%	0%	83%	0%	0%	0%
Accept	100%	0%	0%	0%	0%	0%	0%	0%
Reject	1%	10%	0%	1%	8%	6%	1%	73%

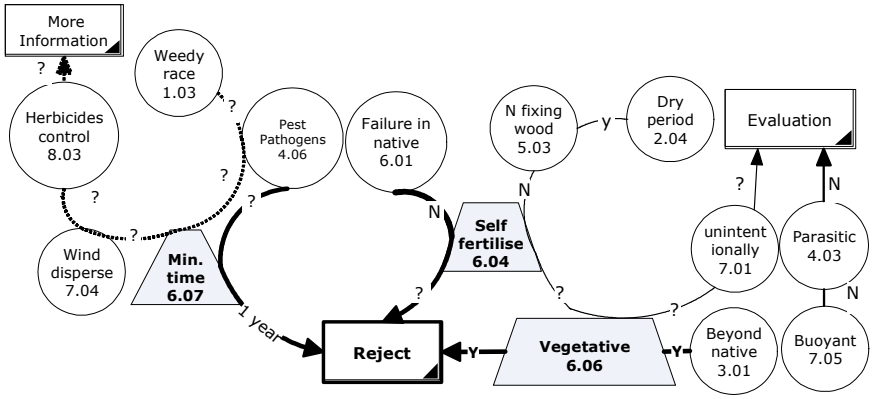


Fig. 3-2 The shortest pathway of selecting key Australia WRA questions using Ant-Miner.

3.3.3. Hawaii/Pacific WRA system

Similarly to the Australia results, Table 3-3 shows a summary of key Hawaii/Pacific WRA questions, detected by TNS. For example, whether the plant is an environmental weed or not accounts for 63% of the decisions in the decision tree (*Prop I* (a_i) value) resulting in a classification of *low risk*. On the other hand, for plant species classified as *high* risk or *evaluation required*, respectively 40% and 14% of decisions in the tree (*Prop I* (a_i) value) are for the question of whether the species is highly domesticated or not (Section 1.01 in Appendix 3-1).

Fig. 3-3 shows the shortest pathway, identified by Ant-Miner, to select the key questions (nodes). Interestingly, the questions (nodes) that were identified to predict each class do not overlap between *reject* (left side of Fig. 3-3) and *low risk* and *evaluation* (right side of Fig. 3-3). This suggests that the Hawaii/Pacific WRA system has a strong structure to make a decision for each class. For example, the *high risk* plants are assessed particularly by whether the plant is *beyond native or not* (3.01), shown as the starter question (node) in Fig. 3-3 (left). If the plant is *introduced outside its native range* (2.05), and is *beyond native* (3.01) and tolerates or benefits from mutilation, cultivation or fire then the plant species is rejected.

Table 3-3 Hawaii/Pacific WRA key questions selected by TNS, with TNS decision proportions for each class shown as percentages (*Prop I* (a_i) value).

Hawaii Pacific	Section 1	Section 3		Section 4				Section 6		Section 7
	Domestication/ cultivation	Weed elsewhere		Undesirable traits				Reproduction		Dispersal mech.
	1.01: highly domesticated	3.03: Agr./hort./ forestry	3.04: Env.weed	4.06: Recog. Pests &pathogens	4.09: shade tolerant	4.11: Climbing /smothering	4.12: Dense thickets	6.02: Viable seed	6.06: Veg. Frag.	7.06: Bird
Eval.	14%	3%	0%	3%	28%	24%	0%	0%	0%	28%
Low	0%	31%	63%	0%	1%	0%	1%	0%	1%	2%
High	40%	37%	0%	0%	0%	4%	1%	16%	1%	2%

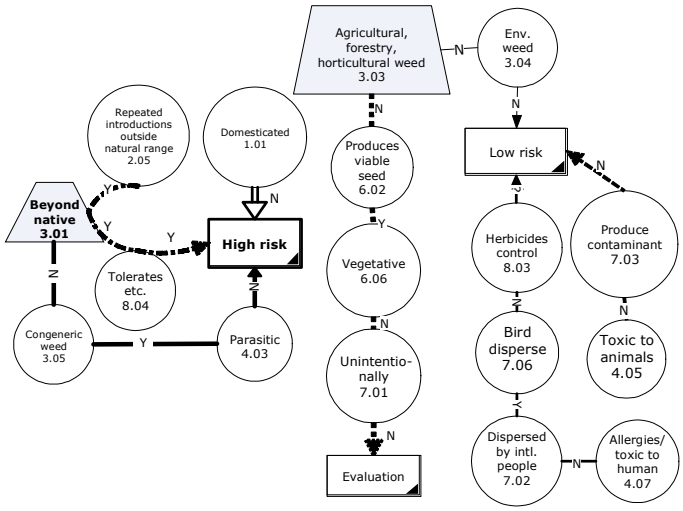


Fig. 3-3 The shortest pathway of selecting key Hawaii/Pacific WRA questions using Ant-Miner.

However, if the plant is not beyond native, but is recognised as *congeneric weed* (3.05) and *parasitic* (4.03), then the plant species is rejected. Also, if the plant is *not domesticated* (1.01), then the plan species is rejected. The *low risk* and *evaluation* classes are commonly assessed, when the weed is not found from agriculture, *horticulture or forestry* (3.03).

3.3.4. Assessment trends of the WRA between TNS and Ant-Miner, and between Australia and Hawaii/Pacific WRA models.

The summary of selected key questions identified by Ant-Miner is demonstrated by a leaf-stem plot for Australia results in Fig. 3-4 (left) and Hawaii/Pacific in Fig. 3-4 (right). Each number in a leaf-stem plot (Fig. 3-4) is summarized from nodes that were identified by Ant-Miner results. The questions that Ant-Miner also selected for the other WRA, i.e. questions selected for both Australia and Hawaii/Pacific WRA, are indicated in bold. A star (*) indicates that the question was selected by TNS for both Australia and Hawaii/Pacific WRA. Note that question numbers appearing more than once in the leaf-stem plot indicate that the

The WRA question		Australia		Hawaii/Pacific
Dometication/cultivation	1	03	1	01*
Climate and distribution	2	04	2	05
Weed elsewhere	3	**01***	3	**01, 01,** 03*, 04*, 05
Undesirable traits	4	**03**, 06	4	**03**, 05, 07
Plant type	5		5	
Reproductoin	6	01, 03, 04, 04, **06, 06,** 07, 07	6	06*
Dispersal mechanisms	7	**01**, 04*, 05*	7	**01**, 02, 03, 06*
Persistence attributes	8	**03**	8	**03**, 04

Fig. 3-4 Summary of questions selected by Ant-Miner for Australia WRA (left) and Hawaii/Pacific (right).

Each number in the leaf-stem plot indicates the WRA question's section. Questions selected by Ant-Miner for both WRA models are shown in bold. Questions selected by TNS for both models are marked with a star (*).

question is used to predict different pathways, e.g., 6.06 (*vegetative*) was used to predict both *reject* and *evaluation* (Fig. 3-2).

Both TNS and Ant-Miner identified *dispersal mechanisms* (Section 7, Appendix 3-1) as common and most frequently used questions for the Australia WRA decision making process, i.e., two star marks for 7.04 (*wind dispersal*) and 7.05 (*buoyant*) in Fig. 3-4 (left). However, Ant-Miner selects the highest number of questions (five questions; 6.01, 6.03, 05, 6.06 and 6.07) from *reproduction* questions (Section 6). For Hawaii/Pacific WRA, both TNS and Ant-Miner identified *weed elsewhere* (Section 3) as common and most frequently used questions, i.e., two star marks for 3.03 (agricultural, *horticulture and forestry*) and 3.04 (*environmental weed*), in Fig. 3-4 (right). Besides, Ant-Miner selected the highest number of questions from *weed elsewhere* (Section 3); four questions (3.01, 3.03, 3.04 and 3.05), shown in Fig. 3-4 (right).

This suggests that improving questions such as those identified by both TNS and Ant-Miner, by setting up more specific and detailed questions, may increase sensitivity and help overall judgment. On the other hand, questions identified as related to the class of *more information* may be removed or have aspects changed to ease answering further, which may help creating the cost and time effective WRA analysis.

Following is a summary of key questions that were identified by TNS and Ant-Miner with respect to each class (details of WRA section in Appendix 3-1).

Australia WRA:

Reject : 3.01 (*Naturalised beyond native*)

: 7.05 (*Buoyant*)

Evaluate : 7.04 (*Adapted to wind dispersal*)

Hawaii/Pacific WRA:

Reject : 1.01 (*Highly domesticated*)
: 3.03 (*Agr. hort. forestry: weed elsewhere*)
Low risk : 3.03 and 3.04 (*Env. weed*)
: 6.06 (*Reproduction by vegetative fragmentation*)
Evaluation : 7.06 (*Dispersal by birds*)

This investigation provides insights into the fundamental structures of the WRA systems. Since the *reject* class dominated the Australia WRA data, the key questions focused on *rejected* species. However, key questions selected in common by both TNS and Ant-Miner were *naturalized beyond native* (3.01) and *buoyant* (7.05) for the high risk plants, and the *wind dispersal adaption* (7.04) question tends to be unknown among plants assessed in the Australia model. On the other hand, the class structure of the Hawaii/Pacific data was more balanced; TNS and Ant-Miner identified common key questions for all three classes (details shown above), but different key questions were selected from the Australia model. For example, the high risk plants for Hawaii/Pacific WRA were assessed based on *highly domesticated* (1.01) and *weed of agriculture/horticulture/forestry* (3.03). It is expected that different region and climate would consider different factors to assess the high risk plants for the country and this investigation helps identifying how they were different by identifying key questions. Overall, TNS and Ant-Miner identified w*eed elsewhere* (Section 3) as an important question for the *high risk* (or *reject*) plant, and *dispersal mechanism* (Section 7) tended to be unknown (as it predicts *evaluation required*). It could be said that improving or even considering removing (if the question is consistently difficult to answer for many species) these question carefully during the decision making process will help the future WRA model. If these questions can be more specific and allow the assessment to be more accurate, the overall classification, assessment process, may be improved.

3.4. Conclusions

The Tree Node Selection method (TNS) and Ant-Miner algorithms were applied to identify the key questions in the weed risk assessment model (WRA). TNS searched for key questions by investigating the C4.5 decision tree whereas Ant-Miner identified the shortest pathway (the most dominant and important pathway) to identify questions as nodes in respect to different plant risks. Despite their differences in the heuristic functions, the TNS and Ant-Miner algorithms selected similar key questions and sections respectively for the Australia and Hawaii/Pacific WRA models. Identifying such key questions provides ideas on how different regions with different climates have different risk assessment and decision process

systems. Generally, for assessing the high risk plant species for the Australian and Hawaii/Pacific systems, both TNS and Ant-Miner identify the questions in the *weed elsewhere* category (Section 3) as key questions, suggesting that the weedy nature of plants is an important factor to make a plant risk decision. For Australia and Hawaii/Pacific WRA, TNS and Ant-Miner identified that the dispersal mechanisms (Section 7) are the most unknown questions (class: *evaluation*). The Australia WRA mainly consisted of *reject* or *evaluation* classes whereas the Hawaii/Pacific WRA data consisted of balanced classes for *high* and *low* risk, and *evaluation*. Hence, TNS and Ant-Miner identified *weed elsewhere* (Section 3) and *reproduction mechanism* (Section 6) as important questions for the low risk plants for the Hawaii/Pacific WRA.

Identifying influential factors from the model helps construction of cost effective biosecurity strategies, with well designed questions to provide reliable decisions through the WRA model. This investigation provides information to target the most important questions and encourage the investigator to answer those as carefully and accurately as possible. This effort may help the decision making process for the plant importation system by improving the accuracy to identify plants through the WRA model.

In addition to the use of TNS, this study also shows that the uncommonly used Ant-Miner can be a useful data mining tool, as it successfully provided important pathways for assessing different risks. At this stage, this investigation did not aim to construct new risk models, but rather to increase knowledge about the existing model. In the future, many more different plant species and data points taken from different regions will be investigated to help improve the WRA model.

From this study, Chapter 7 introduces future work on developing a website project for the WRA model, the WRA Information Database Service (WRA-IDS). The WRA-IDS is proposed to provide a searchable archive of alien plant information provided by many scientists, to help with the WRA model process. It will allow scientists to comment and add their own data, and will incorporate simple statistics on the questions, with online attribute selection tools to help the decision making system for the alien plants.

3.5. References

Daehler CC, Denslow JS, Ansari S, Kuo H (2004) A risk assessment system for screening out invasive pest plants from Hawai'i and other Pacific Islands. Conserv Biol 18:360-368.

Dorigo M, Stützle T (2004) Ant Colony Optimization, MIT Press/Bradford Books, Cambridge, MA.

Fukuda K, Brown J (2007a) Investigation of The Weed Risk Assessment Model Using Data Mining, Intl Conf. of 9th EMAPi9, abstract.

Fukuda K, Brown J (2007b) Classification Rule Extraction by Ant-Miner for Weed Risk Assessment, In Oxley L and Kulasiri D. (eds) MODSIM 07, 2882-2888.

García-Martínez C, Herrera F (2007) A taxonomy and an empirical analysis of multiple objective and colony optimization algorithms for the bi-criteria TSP, Eur J Ope Res 180: 116-148.

Parpinelli RS, Lopes HS, Freitas AA (2002) Data mining with and Ant Colony Optimization algorithm, IEEE Trans Evol Comput 6: 321-332.

Pheloung PC, Williams PA, Halloy SR (1999) A weed risk assessment model for use as a biosecurity tool evaluating plant introductions, J Environ Manag 57: 239-251.

PIER (2007) Institute of Pacific Islands Forestry Pacific Island Ecosystems at Risk (PIER) Plant threats to Pacific ecosystem. Available via http://www.hear.org/pier. Accessed on 15 September 2007.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Rejmanek M (2001) In Groves, RH, Panetta FD and Virtue JG (eds) Chapter 1: What tools do we have to detect invasive plant species?, 3-7. CIRO publishing, Victoria, Australia.

Williamson M (2001) In Groves, RH, Panetta FD and Virtue JG (eds) Chapter 3: Can the impacts of invasive species be predicted?, 20-31. CIRO publishing, Victoria, Australia.

Yeates GW, Williams PA (2001), Influence of three invasive weeds and site factors on soil microfauna in New Zealand, *Pedobiologia*, 45, 367–383.

3.6. Appendices

Appendix 3-1 Weed risk assessment model questions (Pheloung et al. 1999).

Appendix 1

Questions forming the basis of the Weed Risk Assessment model (WRA)

Weed Risk Assessment system question sheet: Answer yes (y) or no (n), or don't know (leave blank), unless otherwise indicated

	Botanical name:		Outcome:	
	Common name:		Score:	
	Family name		Your name:	
		History/Biogeography		
A	1 *Domestication/ cultivation*	1.01	Is the species highly domesticated? If answer is 'no' got to question 2.01	
C		1.02	Has the species become naturalised where grown?	
C		1.03	Does the species have weedy races?	
	2 *Climate and Distribution*	2.01	Species suited to Australian climates (0–low; 1–intermediate; 2–high)	2
		2.02	Quality of climate match data (0–low; 1 intermediate; 2–high)	2
C		2.03	Broad climate suitability (environmental versatility)	
C		2.04	Native or naturalised in regions with extended dry periods	
		2.05	Does the species have a history of repeated introductions outside its natural range?	
C	3 *Weed elsewhere*	3.01	Naturalised beyond native range	
E		3.02	Garden/amenity/disturbance weed	
A		3.03	Weed of agriculture/horticulture/forestry	
E		3.04	Environmental weed	
		3.05	Congeneric weed	
		Biology/Ecology		
A	4 *Undesirable traits*	4.01	Produces spines, thorns or burrs	
C		4.02	Allelopathic	
C		4.03	Parasitic	
A		4.04	Unpalatable to grazing animals	
C		4.05	Toxic to animals	
C		4.06	Host for recognised pests and pathogens	
C		4.07	Causes allergies or is otherwise toxic to humans	
E		4.08	Creates a fire hazard in natural ecosystems	
E		4.09	Is a shade tolerant plant at some stage of its life cycle	
E		4.10	Grows on infertile soils	
E		4.11	Climbing or smothering growth habit	
E		4.12	Forms dense thickets	
E	5 *Plant type*	5.01	Aquatic	
C		5.02	Grass	
E		5.03	Nitrogen fixing woody plant	
C		5.04	Geophyte	
C	6 *Reproduction*	6.01	Evidence of substantial reproductive failure in native habitat	
C		6.02	Produces viable seed.	
C		6.03	Hybridises naturally	
C		6.04	Self-fertilisation	
C		6.05	Requires specialist pollinators	
C		6.06	Reproduction by vegetative propagation	
C		6.07	Minimum generative time (years)	1
A	7 *Dispersal mechanisms*	7.01	Propagules likely to be dispersed unintentionally	
C		7.02	Propagules dispersed intentionally by people	
A		7.03	Propagules likely to disperse as a produce contaminant	
C		7.04	Propagules adapted to wind dispersal	
E		7.05	Propagules buoyant	
E		7.06	Propagules bird dispersed	
C		7.07	Propagules dispersed by other animals (externally)	
C		7.08	Propagules dispersed by other animals (internally)	
C	8 *Persistence attributes*	8.01	Prolific seed production	
A		8.02	Evidence that a persistent propagule bank is formed (>1 yr)	
A		8.03	Well controlled by herbicides	
C		8.04	Tolerates or benefits from mutilation, cultivation or fire	
E		8.05	Effective natural enemies present in Australia	

A = agricultural, E = environmental, C = combined.

Study II. Assessment of the structure of decision trees by TNS and TNS-A for sea container contamination using biosecurity risk profiles

3.7. Introduction

The motivation of this study was to introduce a use of computer algorithms as knowledge discovery tools, to allow flexible and computationally efficient analysis of the unique nature of environmental science data – non-numerical and categorical data that can contain attributes with many unique values, many of which occur only a few times in the dataset – without modifying or removing instances. When the dataset may only have a relatively small number of attributes and instances, it is important to select a method that does not require removing instances, which will further reduce its size. Here, Tree Node Selection (TNS) and an additional developed tool, Tree Node Selection for assessing decision tree structure (TNS-A), are demonstrated to help understanding the sea container contamination pathway for New Zealand Border Biosecurity using risk profiles.

Firstly, the TNS method, developed in Chapter 2, was used to rank factors – attributes involved in the risk profiles of the sea container contamination for New Zealand Biosecurity – by their importance. Previously, TNS was used for attribute selection (Chapter 2 and Chapter 3, Study I) to select a subset of important attributes from many attributes, whereas in this case, the studied data contained only six attributes, and TNS was used to rank them in order of importance, to help identifying important factors for the future biosecurity policy and management process. Since the base algorithm for TNS, the C4.5 algorithm, is often compared with statistical approaches such as logistic regression as an off-the-shelf method for building classification models (Perlich et al. 2003), this study briefly covers the application difficulty and comparison of inspecting the nature of this study data with a statistical approach.

Secondly, the decision tree assessment tool, TNS-A, developed in this chapter, is used to extract further knowledge about the association and relationship between potential factors for the sea container contamination decision making process by assessing the decision tree structure.

3.7.1. Sea container contamination for New Zealand Biosecurity

The spread of exotic pests poses a worldwide threat. The World Trade Organization (WTO) and the Sanitary and Phytosanitary (SPS) Agreement of 1994 announced international rules to protect human, animal, or plant life from risks associated with additives,

contaminants, toxins and disease, and to protect a country from damage caused by the entry, establishment or spread of pests (WTO 1995). The New Zealand Biosecurity Act was enacted in 1993, and the first New Zealand Biosecurity Strategy, developed in 2003, recommended the immediate implementation of several steps: to identify, prioritise and review current and emerging risks, establish national leadership and coordination of pest management, recognise the contribution of science to biosecurity and fund it properly, and ensure decision-making processes take account of risks to the economy, biodiversity, taonga, human health and lifestyle in setting priorities (Biosecurity Council 2003). Historically, New Zealand has protected its agricultural, horticultural and forestry industries (Taylor et al. 2000), but increasing demands from tourism and trade have necessitated the strengthening of border controls and inspection to prohibit the entry of alien pests, weeds and diseases affecting the environment or human health.

Just over 50 years ago, the first shipment of sea containers took place between New Jersey and Texas. Today, sea containers transport approximately 90% of world-wide cargo (Ports of Auckland Ltd 2006). An unintended consequence of world-wide transport is that sea containers may carry exotic organisms to new places (Border Management Group 2003). For New Zealand, with a significant world trade but no land borders, sea containers represent a significant pathway for the potential entry of unwanted organisms. Containers not only bring risks to New Zealand ports, but transport them inland to importers' and exporters' premises. The number of sea containers imported into New Zealand has grown by 54% between 2000-01 and 2005-06. The volume of containerised cargo imported has grown even more, as 40ft-containers are replacing 20ft-containers. Just over half of New Zealand's imported containers arrive at Auckland, with 21% arriving at Tauranga and 9% at Lyttelton.

The results of a New Zealand Ministry of Agriculture and Forestry (MAF) sea container survey in 2001-02 (Border Management Group 2003) and subsequent consultation with stakeholders led to substantial changes in MAF's sea container risk management. A new standard for sea container risk management was implemented on 1 January 2004, and in October of 2004, an electronic sea container risk profiling system was introduced, allowing MAF to electronically select high-risk sea containers for inspection. In July 2006, MAF also implemented an international standard for wood packaging material, known as ISPM-15 (ISPM 2002) designed to reduce the world-wide spread of timber pests and diseases through wood packaging.

MAF develops container risk profiles based on import and inspection data. It is significantly important for MAF to understand the system of the sea container contamination

pathway in order to manage and mitigate further contamination. These profiles are made up of a series of criteria that identify containers posing a higher than average probability of being contaminated. Selection of appropriate criteria is important to ensure that inspection resources are used effectively and efficiently. Brookmeyer (2005) reported the significant role of statisticians in the context of bioterrorism in USA that: 1) biosecurity policy decisions must be based on the best available science, 2) statisticians have much to contribute and should be actively working with multidisciplinary teams of experts and 3) statisticians can make a difference.

Results of detected important potential risk factors and their relationship to the pathway of sea container contamination in this study were explained and discussed with MAF in order to encourage more use of data mining as a knowledge discovery tool. Currently, MAF is continuing to investigate various techniques for evaluating criteria, including data mining. Knowledge gained from this study was also discussed with MAF to construct an early warning system for detecting future sea container contamination risk, even before the ships enter the country (discussed in Chapter 7).

3.7.2. Ranking important risk factors by TNS

Firstly, the Tree Node Selection (TNS) method, developed in Chapter 2, was used to rank important factors using sea container contamination risk profiles to understand the decision making process for the sea container contamination pathway. The study data consisted of six risk profiles of sea container contamination: *wood type* (WT), *container type* (CT), *port* (P), *vessel last region* (VL), *port of loading region* (POL), and *content region* (CR), and whether contamination was detected or not by MAF (the class; *yes* or *no*). These six attributes were the most conveniently available information on containers prior to arrival and inspection in New Zealand; the risk profile of the sea container data involves other variables, such as *container contents*, which consists of multiple values, for example a single container might have its contents listed as *carpet*, *flour*, *CDs*. This is unfortunately not useful at this point as mining this type of data requires more training examples than are currently available.

The dataset contained about 1,400 instances (recorded over two years in 2001-2002), all of which consisted of non-numerical values, with many attributes containing values that occur infrequently, and at times once only (a summary of instances is shown in Appendix 3-2). All attributes were described by discrete text values, for example, the values of the *port* attribute are Auckland, Lyttelton, Mt. Maunganui, and Napier. The study data contained a total of 152 unique non-numerical values among 6 attributes, e.g., attribute *port of loading* (*POL*)

contained 113 different values (Appendix 3-2). Ideally, a flexible method will be able to handle the unique nature of the data without modifying or removing attributes or instances.

Some statistical methods, such as logistic regression, are often compared with the C4.5 algorithm (a base algorithm for TNS), since they assess the quality of rankings based on class membership probabilities and work on binary data, e.g., Lim et al. (2000), Perlich et al. (2003). Some other statistical methods, e.g., principal component analysis and best subsets regression, are generally applied on numerical data, thus may not be directly applicable for the study data, since both the inputs and outputs of the study data were completely categorical. Perlich et al. (2003) carefully examined the performance of (binary) logistic regression and the C4.5 algorithm using learning curves analysis and concluded that while logistic regression does not generally outperform tree induction, logistic regression was better for smaller training sets (fewer than approximately 1,000 instances, based on experimenting on 36 benchmark datasets with 320 to 1 million instances, with a median size of 12,800). He concluded that tree induction was better for larger data sets.

Logistic regression requires the test set to contain only those nominal values that have been seen previously in the training set (Perlich et al. 2003). This means that if the training sample does not contain the value "Osaka" for the attribute port, for example, logistic regression cannot estimate a parameter for this variable and will produce an error message and stop execution when a test example with *port* = "Osaka" appears. In this study this would mean 113 instances would be lost for attribute *port* instantly (details in Appendix 3-2), which may not be ideal, if the investigator aimed to obtain the maximum possible knowledge from the data.

It is also important to consider the computation time. Different statistical packages process the data differently, but, for example, logistic regression often takes an excessively long time to run, even on only moderately large data sets (Perlich et al. 2003) and some packages, e.g., Minitab 15, cannot even execute it on very large data. For example, if the best subsets regression was applied to this study data (a total of 152 unique non-numerical values, as previously mentioned), finding the best possible submodel or attribute set would require considering more than 2 million different possibilities just to evaluate all four-attribute subsets ($^{152}C_4$). In comparison, C4.5 can handle this problem by splitting the example probabilistically and sending weighted (partial) examples to descendant nodes (Perlich et al. 2003, details in Quinlan 1993). It could be argued that such unique instances, for example the 113 single values for attribute *port* may not be relevant or important for the investigation, but removing an entire instance consequently removes information from other attributes.

Similarly, it is important to keep attribute *port* for the analysis, to try to understand the potential risk factors for the entire sea container contamination pathway, especially when there are not too many available attributes to investigate. If the attribute *port* is irrelevant, the C4.5 algorithm would answer this question by ranking the attribute lowest anyway.

Here, possible advantages of applying a tree induction algorithm and TNS for the study data or similar to such data would be: 1) most tree induction algorithms including TNS are flexible enough to handle data sets of various natures, e.g., discrete, continuous, missing, binary and non-linear domains; 2) computational efficiency, i.e., TNS ranks the attributes in on average 2 seconds, excluding evaluating the selected attribute subsets (from testing on less than 300 attributes and less than 50,000 instances, shown in Table 2-6 in Chapter 2); 3) TNS was found to provide the most consistent performance among other attribute selection methods (Chapter 2); 4) TNS provides knowledge about attributes based on the decision making process, which is suitable to understand the sea container contamination decision making process, and 5) when the dataset is small, such as the study data, which only has 1400 instances, it is generally recommended to apply a simple and well-known data mining algorithm (Spate et al. 2006), perhaps such as the C4.5 algorithm. Generally, data mining algorithms were designed for very large data sets, e.g., millions of attributes for text recognition, thus the TNS method, based on the C4.5 algorithm, can be applied and expected to provide realistic results for both small and very large data sets.

3.7.3. Decision tree assessment tool, TNS-A

Additionally, this chapter developed the decision tree assessment tool, TNS-A, based on the TNS method, to add extra knowledge about attributes and their relationship to the *class* or decision, by assessing the decision tree structure that was generated with the best subset of attributes (assessed from TNS).

Construction of decision trees is not only useful for prediction or classification purposes; they describe the decision process in a readable, comprehensible manner, which can be used for knowledge discovery. Manual visualisation or interpretation of the decision tree can be difficult; especially when the decision tree is large, it can be almost impossible to interpret or summarise. If the total number of input attributes is small, such as six in this case, it is easy to visualise the decision tree structure as an interaction of nodes. However, quantifying or identifying the decision structures by connecting correctly (or incorrectly) classified instances to the nodes or trying to understand the relationships between nodes, and between nodes and decisions, in the decision tree, can be complicated unless the tree size is very small, e.g., only a few branches.

Generally, the assessment of decision trees is carried out by developing visualization tools to obtain knowledge about the decision tree by displaying the structure of the entire decision tree, e.g., Ankerst et al. (2000). Barlow and Neville (2001a,b), Teoh and Ma (2003). For example, Ankerst et al. (2000) turned the decision tree structures into pixel-oriented visualization techniques to map each attribute value of each data object to one coloured pixel and to represent the values belonging to different attributes, Barlow and Neville (2001a,b) drew the smallest organization chart of the entire decision tree to provide useful information about the tree and Teoh and Ma (2003) displayed each node in the decision tree as a visual projection of the data. As a result, the obtained information was used to improve the classification accuracy (Teoh and Ma 2003) or tree size (Ankerst et al. 2000) of decision tree construction as like TNS and attribute selection methods.

In comparison to visualization of the tree, in order to gain knowledge about the decision tree, TNS-A, was developed in this chapter. TNS-A assesses the decision tree by counting the number of instances that are classified by paths passing through each node (attribute) and between nodes to the leaf node (class) to identify the relationship or associations between attributes, and between attributes and classes. In other words, this assessment helps understanding the association of each factor, between pairs of factors, and between factors and decisions about sea container contamination. Note that this study focuses on the relationship between attributes and the contamination decision. In fact, the specific attribute values (edges) are not described in results to avoid sensitive issues. Lastly, results of TNS and TNS-A, attribute rankings, association of each attribute, and attributes and decision, are represented in matrix form to make interpretation simpler and easier.

In this study, the data were investigated for the full data set (over two years) and four seasons to help enhance different risk profiles among seasons, since interestingly, sea container risk profiles have some seasonal effects, e.g., there is an increased level of importation from various regions in the month of Christmas. To examine the maximum aspects of data; 1) the original classification accuracy and decision tree structures of C4.5 (without TNS) and naïve Bayes classifier were examined, 2) selected attributes by TNS and the assessment of decision trees by TNS-A were examined, and 3) the effect of removing the least important attribute selected by TNS on the classification accuracy of the C4.5 classifier was demonstrated.

Table 3-4 Summary profiles of the sea container data sets (in number of instances).

Contamination records (the class)	Full[a]	Mean[b]	SD[b]	Container type (CT)	Full[a]	Mean[b]	SD[b]	Port (P)	Full[a]	Mean[b]	SD[b]
No	837	209.3	19.1	Bulk	9	2.3	1.7	Auckland	962	240.5	22.1
Yes	564	141.0	15.6	Flat rack	22	5.5	1.7	Lyttelton	195	48.8	6.9
Total	1401			General	1234	308.5	22.8	Mt. Maunganui	195	48.8	5.9
Wood type (WT)	Full[a]	Mean[b]	SD[b]	Hazardous	7	2.3	0.6	Napier	49	12.3	1.3
N/A[c]	734	183.5	23.8	N/A[c]	42	10.5	8.6				
Packaging	398	99.5	6.0	Open	20	5.0	1.8				
Packaging and dunnage	154	38.5	4.7	Reefer	66	16.5	5.3				
Dunnage	115	28.8	4.6	Tank	1	0.3	0.5				
Vessel last region (VL)	Full[a]	Mean[b]	SD[b]	POL region (POL)	Full[a]	Mean[b]	SD[b]	Content region	Full[a]	Mean[b]	SD[b]
Asia and Middle East	222	55.5	10.3	Africa	10	2.5	1.3	Africa	20	5.0	2.2
Australia	740	185.0	24.4	Asia and Middle East	251	62.8	8.5	Asia and Middle	282	70.5	11.6
Central *S* America	8	2.0	1.8	Australia	464	116.0	14.9	Australia	418	104.5	16.3
EU/Scandinavia	12	4.0	6.9	Central S America	7	1.8	0.5	Central S America	10	2.5	0.6
Japan	22	5.5	1.3	EU/Scandinavia	155	38.8	7.4	EU/Scandinavia	175	43.8	3.4
New Zealand	11	2.8	2.1	Japan	47	11.8	3.8	Japan	42	10.5	3.8
N/A[c]	18	6.0	9.5	N/A[c]	25	6.3	2.9	New Zealand	1	0.3	0.5
North America	118	29.5	2.5	North America	108	27.0	2.7	N/A[c]	102	25.5	4.9
Pacific Islands	39	9.8	7.8	Pacific Islands	7	1.8	1.0	North America	110	27.5	1.3
South East Asia	198	49.5	5.8	South East Asia	315	78.8	10.4	Pacific Islands	7	1.8	1.0
Unknown	13	3.3	5.3	Unknown	12	3.0	3.2	South East Asia	164	41.0	1.4
								Unknown	70	17.5	7.4

[a] Full data sets (all four seasons).
[b]. Mean and standard deviation (SD) is calculated from the four seasonal data sets.
[c] .N/A indicates no data entry. Unknown indicates the value was entered as unknown.

3.8. Data and method

The study data is introduced below, followed by the TNS-A algorithm. While the TNS algorithm was explained in Chapter 2, this section introduced how the TNS-A algorithm was extended from TNS.

3.8.1. Data set

The sea container data set, provided by MAF, represents a subset of the data collected during the 2001-02 survey (Border Management Group 2002). Table 3-4 shows a detailed data set profile, indicating numbers of instances for each attribute value. Note that *N/A* and *Unknown* in Table 3-4 indicate the data entry is *blank* or *unknown*, respectively. Detailed information of instances for each attribute is summarised and shown in Appendix 3-2.

Six input attributes ($A = 6$) are taken from July 2001 to December 2002 ($I = 1401$ instances): *wood type* (WT), *container type* (CT), *port of loading* (POL), content region (CR), *vessel last region* (VL) and *port* (P). The container contamination record (*no* or *yes*) is used as the class for decision tree construction. Five data sets of different lengths are prepared: the full data set ($I = 1401$) and four shorter sets, each of which contains a single season from the full data set: summer ($I = 319$), covering December, January, and February; autumn ($I = 370$),

March, April, and May; winter (I = 329), June, July, and August; and spring (I = 383), September, October, and November. Table 3-4 shows the mean and standard deviation (SD) of the instances, calculated from the four seasonal data sets to compare with the full data set. The distribution of the class (*no* or *yes*) in the full data set is skewed towards the class *no*, which has 837 instances (59.7%), whereas the class *yes* has 564 instances (40.3%). Generally, larger SD values, e.g., *Australia* for VL, in Table 3-4 suggest that the frequencies of particular attribute values vary between seasons.

3.8.2. Tree Node Selection for assessing decision tree structure, TNS-A

The TNS-A method was developed based on the concept of TNS (details of the TNS algorithm are shown in Section 2.2.3.1, Chapter 2). TNS assesses each node in the tree by counting the number of instances that are classified by a path passing through the node, and ranking the overall contribution for each attribute by the sum of such instance counts for all nodes labelled with the given attribute. TNS results were used to rank factors for the decision support system in this study, whereas TNS-A identifies three different aspects by assessing the decision tree, generated with the best subset of attributes (identified by TNS).

Firstly, TNS-A calculates the total number of instances that is classified to each class, indicating a summary of how the generated decision tree classified all instances to each decision; contamination *yes* or *no* (the algorithm is in Section 3.8.2.1).

Secondly, TNS-A assesses each node in the tree with each class by counting the number of instances that are classified by a path passing through the node and the class, and ranking the overall contribution for the attribute and the class by the sum of such instance counts between all nodes labelled with the given attribute and all classes labelled with the given class. For example, if there is a higher count between attribute *wood type* and the positive contamination risk (*yes*) than between attribute *port* and the positive contamination risk (*yes*), this indicates that the the *wood type* has a higher association than *port* with positive contamination risk (algorithm shown in Section 3.8.2.3).

Thirdly, TNS-A assesses a pair of attributes, one node ($v_{i,}$) to another node (v_j), in the tree by counting the number of instances that are classified by a path passing between the two nodes ($v_{i,}$, v_j), and ranking the overall contribution for two nodes by the sum of such instance counts between all nodes labelled with the given v_i and all nodes labelled with the given v_j. For example, if higher counts between attribute *wood type* (v_i) and *port* (v_j) than between attribute *wood type* (v_i^*) and *container type* (v_j^*) are detected, this indicates that the association of attributes *wood type* and *port* is higher than *wood type* and *container type* (algorithm shown in Section 3.8.2.4).

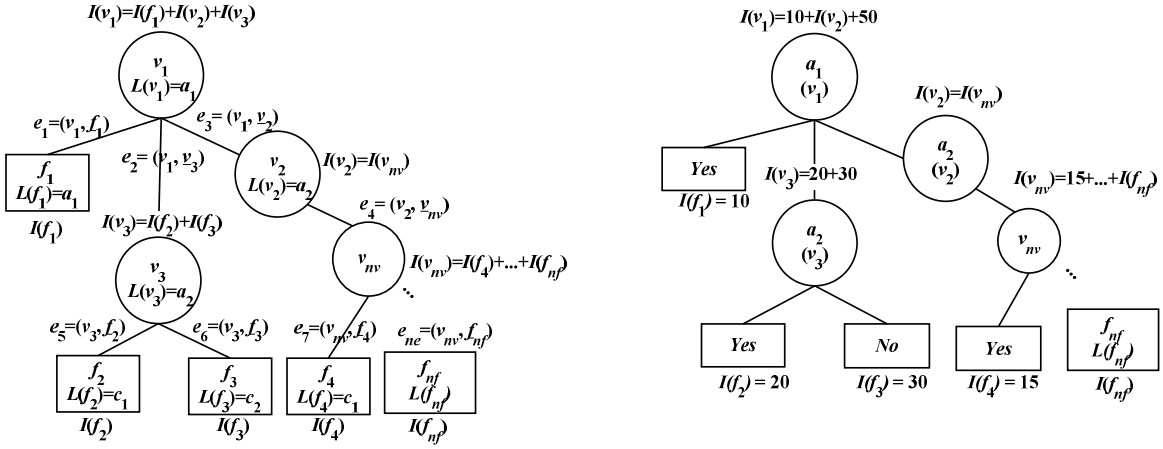


Fig. 3-5 Description of Tree Node Selection process (left) and an example of the decision tree (right), taken from Fig. 2-2 in Chapter 2.

The three assessments that are part of the TNS-A algorithms are described separately: classes, between nodes and classes, and between nodes. Section 2.2.3.1: Tree Node Selection method in Chapter 2 described the concept of TNS in detail, thus, the following section briefly shows the notations for TNS-A.

3.8.2.1. Assessment of predicted class

Let $T = (V, F, E, Lv, Lf)$ be a generated decision tree (Fig. 3-5). The nodes are represented as $V(T) = \{v_1, \ldots, v_{nv}\}$, where nv is the total number of nodes in the decision tree T (excluding leaf nodes). Let A be a set of input attributes where $A = \{a_1, \ldots, a_{na}\}$ and na is the number of attributes. The labels corresponding to the nodes in $V(T)$ are represented as $Lv = \{L(v_1), \ldots, L(v_{nv})\}$ and $L(v_i) \in A \; \forall \; v_i \in V(T)$, where $L(v_i)$ is the label for node v_i.

Similarly, the leaf nodes are represented as $F(T) = \{f_1, \ldots, f_{nf}\}$, where nf is the number of leaf nodes. Hence, the size of the decision tree (T) is $nv + nf$. Let C be a set of classes where $C = \{c_1, \ldots, c_{nc}\}$ and nc is the number of classes. The labels corresponding to the leaf nodes $F(T)$ are represented as $Lf = \{L(f_1), \ldots, L(f_{nf})\}$, and $L(f_i) \in C \; \forall \; f_i \in$ F, where $L(f_i)$ is the label for leaf node f_i. For example, if there are two input classes ($nc = 2$; class c_1 for *yes* and c_2 for *no*) and four leaf nodes were created as $F(T) = \{f_1, f_2, f_3, f_4\}$, the corresponding labels for $F(T)$ might be $L(f_i) = \{c_1, c_1, c_2, c_1\}$, which indicates that leaves f_1, f_2 and f_4 are labelled with the class *yes* (c_1), and f_3 is labelled with the class *no* (c_2) shown in Fig. 3-5 (right) as an example.

Connections between pairs of nodes (including leaf nodes) are represented by edges, $E(T) = \{e_1, \ldots, e_{ne}\}$ where ne is the number of edges in T. An edge e_i between two nodes (v_j and v_k) is defined as $e_i = (v_j, v_k) \mid v_j, v_k \in V(T)$, and an edge e_i between a node v_j and leaf node f_k is defined as $e_i = (v_j, f_k) \mid v_j \in V(T), f_k \in F(T)$.

Let I be the total number of correctly classified instances at a node or leaf node, such that $I(f_i)$ represents the number of correctly classified instanced at leaf node f_i, and $I(v_i)$ is defined recursively (equation 2-1 in Chapter 2).

3.8.2.2. Class assessment

Each class c_i is ranked by the total number of instances classified as c_i, calculated from the sum of the instances at leaf nodes labelled with class c_i (labelled as $L(f_j) = c_i$),

$$I(c_i) = \Sigma\, I(f_j)\ \forall\, f_j \mid L(f_j) = c_i \in C. \tag{3-6}$$

This investigation helps assessing the predicted class distribution in T.

3.8.2.3. Relationship between attributes and classes

The relationship between an attribute and a class (*no* and *yes*) is assessed by counting instances between nodes labelled with the attribute (a_i) and leaf nodes labelled with the class (c_j). Let $I(a_{i,}, c_j)$ be the total number of instances that are classified by an edge between nodes labelled with attribute a_i and leaf nodes labelled with class c_j,

$$I(a_i, c_j) = \Sigma\, I(f_l)\ \forall\ v_k, f_l \mid (v_k, f_l) \in E, \mathrm{L}(v_k) = a_i, L(f_l) = c_j. \tag{3-7}$$

Higher values of $I(a_{i,}, c_j)$ indicate that the specific attribute connects to the leaf node (class) more directly compared to other attributes.

3.8.2.4. Relationship between attributes

The relationship between a pair of attributes is assessed by counting instances that are classified by leaf nodes below an edge linking the two attributes in the pair.

Let $I(a_{i,}, a_j)$ be the total number of such instances for attributes a_i and a_j,

$$I(a_{i,}, a_j) = \Sigma\, I(v_l)\ \forall\ v_k, v_l \mid (v_k, v_l) \in E, L(v_k) = a_i, L(v_l) = a_j \tag{3-8}$$

A higher $I(v_{i,}, v_j)$ value indicates that more instances are classified by a path passing through the specific pair of adjacent attributes compared with other attribute pairs. However, note that this does not assess pairs of nodes that are not connected by an edge in the decision tree.

3.8.3. Representation of TNS and TNS-A results

To allow comparison of TNS and TNS-A results, each instance count is also calculated for the correctly classified proportion of instances, *Prop I correct* (in %). The decision tree outputs of WEKA show the total number of classified instances at each leaf node, $I(f_i)$, and incorrectly classified instances at the leaf node, $I(f_i)$ *incorrect*. For example, the proportion of correctly classified instances for the class, *Prop* (f_i) *correct*, is calculated from

$$Prop\ (f_i)\ correct = \frac{I(f_i) - I(f_i)\ incorrect}{I(f_i)} \times 100\,. \tag{3-9}$$

Similarly, the proportion of correctly classified instances for between attributes and classes $I(a_{i,}, c_j)$, and between attributes $I(a_{i,}, a_j)$ were calculated, *Prop* $(a_{i,}, c_j)$ *correct* and *Prop* $(a_{i,}, a_j)$ *correct*, respectively. It would be interesting to investigate incorrectly classified instances, as WEKA provides information on what and which sea container contamination factors tend to be misclassified. However, to demonstrate TNS and TNS-A in this chapter, all results were examined for only correctly classified instances.

For the purpose of representation, TNS and TNS-A results were summarised by drawing a matrix, shown in Fig. 3-6. This study divides each training dataset into three partitions, i.e., 3-fold cross validation is performed (details in the next section). Hence, each cell in matrix shows a sum of instance counts and an overall *Prop I correct* (in %) from three decision trees, to represent generalised results over the whole training dataset. Each section in the matrix (in Fig. 3-6) is summarised as follows;

- **Section I: Class assessment.** Overall assessment of predicted class (*no* and *yes*) proportions.
- **Section II: Ranking attributes.** Identifying important attributes in the decision system, assessed by counting frequencies of instances classified via each attribute. This represents TNS results.
- **Section III: Relationship between attributes and classes.** Identifying attributes that have a strong relationship between with a class (*no* and *yes*) by counting instances between attributes and classes.

  Section IV: Relationship between attributes. Identifying pairs of attributes that associate directly, and describing the strength of their involvement in decision-making. This is assessed by counting numbers of instances that use pairs of adjacent vertices.

3.8.4. Data preparation and application of TNS and TNS-A

For all experiments, the J4.8 classifier from WEKA 3.4.12 (Witten and Frank 2005) based on the C4.5 algorithm (default settings) was used to generate pruned decision trees from each

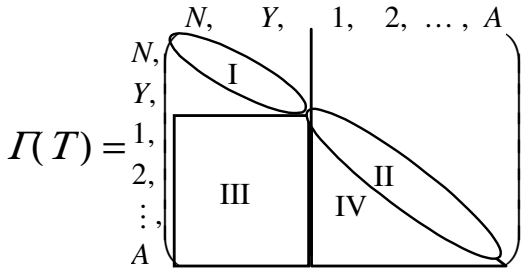


Fig. 3-6 Matrix to represent TNS and TNS-A results.

training data set, thus TNS-P was used (discussed in Chapter 2). All results were then compared between the full and seasonal data sets to identify important factors and their relationships.

Firstly, the original proportions of the classes (*no* and *yes*) and classification accuracy of applying the naïve Bayes and the C4.5 classifier on the original data (without TNS) using 10-fold cross validation are reported. Results from this analysis were only used as supplemental information for the following main investigations for TNS and TNS-A.

Secondly, for TNS and TNS-A applications, training and test data sets were created randomly from the full-length data set and each seasonal data set by selecting two thirds and one third of an unknown data set, respectively, naming the sets Test 1, Test 2 and Test 3 (3-fold cross validation). The same respective proportions of class values, i.e., the proportion of containers classified as *no* or *yes*, were kept to produce training and test data sets to mitigate any bias. Note that TNS in Chapter 2 used a 10-fold cross validation method (details in Section 2.2.2), but the use of two subsets of data to create the training set is also a generally practiced as long as it is tested on an unseen subset of data (Freitas 2002).

The following two experiments were carried out for TNS and TNS-A;

Experiment 1: Three (pruned) decision trees were generated from each training set (two thirds of the data) to produce inputs for TNS and TNS-A. Overall counts of correctly classified instances (and its proportion in %) over three decision trees were reported by drawing the confusing matrix to represent the generalised TNS and TNS-A results.

Experiment 2: Classification accuracy of each (original) decision tree, generated with all attributes, was reported by testing on the unseen third of the data. Several new decision trees (pruned) were constructed by removing attributes selected by TNS one by one, starting from the least important attribute, until only the single most important attribute remained. The most important attribute is ranked 1, and so on. This repeats until the final decision tree is constructed with only one (last) attribute.

The first experiment aimed to provide information about attributes, and the relationships between attributes, and between attributes and decisions, for the sea container contamination decision making system. The second experiment demonstrated how selection of predictive attributes helped constructing an improved decision tree even though the resources were limited to only six attributes.

3.9. Results and discussions

3.9.1. The original C4.5 decision tree and naïve Bayes classifiers

Table 3-5 shows the proportion of the original class (*no*/*yes* in %) and the classification accuracies for the original C4.5 decision tree and naïve Bayes classifiers using 10-fold cross validation (in %) for all sets of data. The overall mean and standard deviation (SD) values of the original class proportions, C4.5 and naïve Bayes were calculated over five sets of data (four seasons and full data). This pre-investigation was helpful to provide general knowledge about the data set and provide some ideas on how data responds to different learning schemes: tree induction for C4.5 and simple probability for naïve Bayes.

Generally, C4.5 shows the higher overall mean the classification accuracy (64.1±3.6%) than the naïve Bayes classifier (63.3±1.8%), whereas the original class proportions for *no* is 59.9±2.1% and *yes* is 40.1±2.1. Thus, it could be said that the results obtained from C4.5 (and naïve Bayes) are at least better than guessing "the container is not contaminated".

However, an interesting point is that separating out seasons from the full data improved the classification accuracy of the decision tree algorithm differently among different methods. The classification accuracy of C4.5 on the full data (65.4%) is higher than the original class proportion (60.0% for *no*), and is similar to the naïve Bayes classifier (65.0%), whereas the NZ winter model has the most improved classification accuracy (66.6%) using the decision tree algorithm; 9.8% higher than the original class proportion (56.8% for *no*). The NZ spring classification accuracy (66.6%) is 3.4% higher than the naïve Bayes classifier (63.2%). The decision tree and naïve Bayes classifiers are found to be comparable for the NZ autumn data, having the same classification accuracy of 63.8%, but better than the original class proportion (60.0% for *no*). However, the NZ summer data shows a classification accuracy using the decision tree classifier (57.9%) that is lower than the original class proportion (62.7% for *no*) and the naïve Bayes classifier (60.2%).

Table 3-5 Summary of original class proportions and classification accuracies (in %) from the C4.5 and naïve Bayes classifiers using 10-cross validation (*n*=10).

NZ season	Summer	Autumn	Winter	Spring	Full data	Overall mean ±SD
Original proportion of the class (*no*/*yes*)	62.7 / 37.3	60.0 / 40.0	56.8 / 43.2	60.0 / 40.0	60.0 / 40.0	59.1±2.1 (*no*) 40.1±2.1 (*yes*)
C4.5	57.9	63.8	66.6	66.6	65.4	64.1±3.6
naïve Bayes	60.2	63.8	64.1	63.2	65.0	63.3±1.8

Despite analysing the unbalanced class data with C4.5, and the varying suitability of the decision tree algorithm between seasons, the application of the decision tree classifier is advantageous, since it provides a decision-making pathway for the sea container contamination profile that can be explored further with the TNS method.

3.9.2. Knowledge discovery for the sea container contamination factors using TNS and TNS-A

The matrix, with overall outputs of TNS and TNS-A, is shown in Fig. 3-7. Each number in Fig. 3-7 represents the overall (total) counts and proportions (in %) of correctly classified instances over three generated decision trees respectively for the full data and for each season. Further, Table 3-6 is made to show a summary interpretation of the TNS and TNS-A output for all sections in the matrix (Fig. 3-7). The numbers in Table 3-6 represent the overall proportions of correctly classified instances (*accuracy* in %, Fig. 3-7) for decision trees for the full and seasonal data sets. The proportion of instances correctly classified by a path passing through an attribute is shown under a rank from 1 to 6 by decreasing order of frequency of use in the decision tree (counts of instance in Fig. 3-7).

In order to generalise the result for discussions, the overall mean and standard deviation (SD) for each class, *no* and *yes*, were calculated from the full and four seasonal data for Section I. Similarly, the overall mean and standard deviation (SD) values were calculated for Section I to IV for each ranked attribute from all data. Overall, the most important attribute was selected from the most frequently detected attribute in the same rank through all data for Section II to IV.

Fig. 3-7 Overall outputs of TNS and TNS-A based on the total counts of classified instances over three decision trees via three training sets.

Total number of instances that were classified is shown as *n* (left) and a proportion of correctly classified instances, *Prop I correct*, is shown as % (right). Blue, orange, grey and white coloured cells describe each section from Section I to IV.

Correctly classified instances		No (*no*)		Yes (*yes*)		Wood type (WT)		POL region (POL)		Container type (CT)		Vessel last region (VL)		Port (P)		Content region (CR)	
		n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
No	Summer	215	71														
	Autumn	391	70														
	Winter	272	74														
	Spring	356	72														
	Full data	1256	72														
Yes	Summer			69	57												
	Autumn			125	70												
	Winter			186	65												
	Spring			168	63												
	Full data			596	63												
Wood type	Summer	215	71	69	57	284	67										
	Autumn	326	69	58	69	516	70										
	Winter	260	74	138	62	458	70										
	Spring	332	71	86	61	524	68										
	Full data	1083	73	121	62	1852	69										
POL region	Summer																
	Autumn																
	Winter																
	Spring																
	Full data	7	88	16	100			23	100								
Container type	Summer																
	Autumn																
	Winter	3	100	44	71	47	72			47	72						
	Spring	14	74	75	64	89	65			89	65						
	Full data	32	76	344	59	293	60			376	60						
Vessel last region	Summer																
	Autumn	22	71	40	66	44	63					62	67				
	Winter	9	60	4	80	13	65					13	65				
	Spring																
	Full data	89	64	55	75	238	67	11	92	83	63	238	67				
Port	Summer																
	Autumn	21	78	4	100	72	77					5	63	72	77		
	Winter																
	Spring																
	Full data																
Content region	Summer																
	Autumn	22	76	23	79	16	89					13	93	42	78	58	81
	Winter																
	Spring	10	83	7	78	17	81									17	81
	Full data	45	68	60	71	117	71	12	86							117	71

Table 3-6 Summary results of the matrix attribute selection shown by the proportion of correctly classification accuracy (in %).

Matrix	**Section I**		**Section II**					**Section III**										**Section IV**					
Relationship	Class		Ranking of attribute					No vs Attribute					Yes vs Attribute					Attribute vs Attribute					
Rank*			Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
Summer	71	57	WT					WT					WT										
			67					71					57										
Autumn	70	70	WT	P	VL	CR		WT	VL	= CR[b]	P		WT	VL	CR	P		WT&P	WT&VL	P&CR	WT&CR	VL&CR	VL&P
			70	77	67	81		69	71	76	78		69	66	79	100		77	63	78	89	93	63
Winter	74	65	WT	CT	VL			WT	VL	CT			WT	CT	VL			WT&CT	WT&VL				
			70	72	65			74	60	100			62	71	80			72	65				
Spring	72	63	WT	CT	CR			WT	CT	CR			WT	CT	CR			WT&CT	WT&CR				
			68	65	81			71	74	83			61	64	78			65	81				
Full data	72	63	WT	CT	VL	CR	POL	WT	VL	CR	CT	POL	CT	WT	CR	VL	POL	WT&CT	WT&VL	WT&CR	CT&VL	POL&CR	POL&VL
			69	60	67	71	100	73	64	68	76	88	59	62	71	75	100	60	67	71	63	86	92
Overall results for full and all seasonal data[a]																							
Attribute	No	Yes	WT	CT	VL	CR		WT	VL	CR	CT		WT	CT	VL	CR		WT&CT	WT&VL	WT&CR			
Mean	71.8	63.6	68.8	65.7	66.3	77.7		71.6	65	75.7	83.3		62.2	64.7	73.7	76		65.7	65	80.3			
SD	1.5	4.7	1.3	6	1.2	5.8		1.9	5.6	7.5	14.5		4.3	6	7.1	4.4		6	2	9			

* Most to least frequently used attributes are ranked from 1 to 6 based on results from Section II in Fig. 3-7.

[a] Mean and standard deviation (SD) are calculated from the full and seasonal data for Section I and from each attribute that is used more than once in any decision pathway for Section II to IV.

[b] VL (vessel last region) and CR (content region) have the same number of instances.

3.9.3. Section I: Assessment of predicted class proportions.

The overall mean classification accuracy for predicting the class *no* is much higher (71.8%) than the class *yes* (63.6%), shown in the lower part of Fig. 3-7, Section I. This may be due to the skewed original proportion of class *no* (about 60% in Table 3-5), which may have helped the prediction of *no*. This result suggests the following scenario; when unknown sea containers are classified via the generated decision trees from the provided data set, 71.8% of containers are classified as uncontaminated (class *no*) correctly, whereas 28.2% are classified incorrectly, i.e., they are actually contaminated; a false negative error. On the other hand, 63.6% of the containers are classified correctly as contaminated (the class *yes*), when 36.4% of them are actually not contaminated (a false positive error). The former type of error should be minimised, because misclassifying contaminated containers as uncontaminated has serious implications from a biosecurity point of view.

However, the lower classification accuracy from the latter suggests that the model shows a conservative judgment towards contamination risk – the error is higher for classifying containers as contaminated when they are in fact not contaminated. To increase knowledge about the sea container contamination profile, investigating the misclassified instances will help understand the pattern of such irregular cases.

Generalised results in Section I, Table 3-6, show that the class *yes* has a higher SD value (±4.7) than the class *no* (±1.5), indicating seasonal differences. The highest and lowest classification accuracies are found from the NZ autumn (70%) and summer models (57%), respectively. The NZ summer model has low classification accuracy for the class *yes*, which indicates that 43% of the sea containers arriving in NZ summer could potentially be misclassified as contaminated (class *yes*), when actually uncontaminated (class *no*). The NZ summer data tends to show unique results, which suggests a specific sea container profile may help the analysis. For example, during busier times of the year, e.g., Christmas in NZ summer, with increased sea container volumes, it would be possible for border inspectors to assess containers more conservatively. Hence, adding extra attributes, e.g., social factor, or weighting a particular attribute, e.g., increasing wood type, may help to describe the summer decision pathway better.

3.9.4. Section II: Identifying important attributes.

The most important and frequently used attribute for all models is the *wood type* (at rank 1; generalised result in Section II, Table 3-6) with a reasonable proportion of correctly classified instances (68.8±1.3%). This classification accuracy may be improved further by

adding a specific description to the wood type, e.g., a treatment type. In fact, MAF has recently adopted the ISPM-15 standard (ISPM 2002) for wood packaging, as previously mentioned. If this results in a change to contamination levels, most important factor in the models could change from *wood type*. The next most important and frequently used attributes are the *container type* (CT), followed by the *vessel last region* (VL) and *content region* (CR), shown from general results, the NZ winter, spring and the full models (Table 3-6). Decisions involving the *content region* show the highest correct classification proportion (77.7±5.8), which confirms that the information gained from the content region is consistent enough to help the correct contamination decision pathway. The *port of loading* (POL) is only found from the full data set as a least important attribute (at rank 5 in Table 3-6), which may suggest that the information about POL does not influence the contamination decision pathway. This makes sense, as the port of loading is often a transhipment port, and may not relate to where the container was packed. Interestingly, the *port* (P) was selected as a secondly important attribute for the NZ autumn model, but was not used for any other data set. This may be due to particular items with specific impacts on the contamination decision pathway, which arrive at a specific port in autumn. Alternatively, the temperature variation among New Zealand ports (Auckland is located in the north, and is often warmer in autumn than Lyttelton, located midway down the South Island) perhaps may have associations with the activity of organisms that could be hidden inside the wood or container shipped during spring in the Northern Hemisphere, but further investigation will be required.

3.9.5. Section III: Attributes associated with the class *no*.

Hidden features – identifying which attributes are frequently associated with the specific decision of *no* or *yes* – are highlighted. Generally, the *wood type* is most and the *vessel last region* (except in NZ spring) is next most frequently used for a decision of the class *no* (rank 1 and 2 in Section III for *No* in Table 3-6). The *container region* and *content type* are generally found later in the order of frequently used attributes (at rank 3 and 4 from generalised results). The least frequently used attributes are the *port* and *port of loading* from the NZ autumn (rank 4) and full (rank 5) models respectively.

The NZ spring model shows a unique result with *container type* being the second most important factor after *wood type* for uncontaminated containers. Interestingly, the NZ winter model shows the lowest proportion of correctly classified instances for the *vessel last region* (60%) compared with other seasons (> 71%), but not with the full data set (64%). In fact, *vessel last region* generally has the lowest overall classification accuracy (65.0±5.6 in Table 3-6), indicating further investigation will be required to see why the *vessel last region*

information confuses the decision of *no*. The third most important attribute for the NZ winter model, the *container type*, has perfect classification (100%), although less frequently used attributes tend to have few instances classified, thus it is not easy to obtain the accuracy. Note that results on the NZ summer data set are discussed in *Section I* (since it uses only *wood type*, results are the same for Section I and III).

Overall, the *wood type* is a significant prediction factor for the contamination profile for *no*, providing reasonably high classification accuracy (71.6±1.9). The *vessel last region* and *content region* also appear to be next important. The analysis suggests that the information particularly from *wood type* and *content region* is reliable, as a reasonably high proportion of correctly classified instances is found (71.6±1.9% and 75.7±7.5% respectively). However, the NZ winter model generally should consider with caution the information gained from *vessel last region* as it has the lowest classification accuracy (60%).

3.9.6. Section III: Attributes associated with the class *yes.*

Generally, the *wood type* is the most significant factor for the sea container contamination decision of *yes*, but the *container type* is most important, when the full data set was analysed (rank 1; Section III for *yes* in Table 3-6). Interestingly, the NZ autumn and spring models select the same attributes between the decisions of *no* and *yes* (autumn, WT, VL, CR and P; spring, WT, CT and CR). However, the range of the NZ spring classification accuracy for *yes* (61-78%, Section III for *yes* in Table 3-6) is lower than *no* (71-83% in Section III, Section III for *no* in Table 3-6). In fact, this trend can be seen from Section II, i.e., the full and all seasonal data sets except NZ autumn show higher classification accuracy for *no* than *yes*, but examination from Section III helps identifying which factors are responsible for lowering or improving prediction of the class. For example, the *container type* has a much lower overall proportion of correctly classified instances for the decision *yes* (64.7±6.0%), compared with the decision *no* (83.3±14.5%). Also, the *wood type* has lower classification accuracy for the decision *yes* (62.2%) compared with the decision *no* (71.6%). Interestingly, the NZ autumn model, which did not select the *container type*, has much higher classification accuracy for the class *yes* (70% in Section I) than the other seasons, which did select the *container type*. On the other hand, the decision-making process involving the *content region* shows a high proportion of correctly classified instances (76.0±4.4%), indicating it provides reliable information for the contamination decision *yes*. From here, further investigation will be helpful to find reasons for lowering classification accuracy, when involving both *wood type* and *container type*.

3.9.7. Section IV: Relationship between attributes.

The TNS-A method successfully detects the relationship between attributes, such that *wood type* is the most important attribute; it significantly associates with other attributes to form a single decision pathway (between two attributes). The most frequent decision pathway is *wood type* and *port* for the autumn data, and *wood type* and *container type* for winter, spring and the full data set (Section IV in Table 3-6). Since the *content region* provides a reasonably high proportion of correctly classified instances, the pair of *wood type* and *content region* shows the highest classification accuracy (80.3±9.0%). The lowest classification accuracy is found from coupling *wood type* with *vessel last region* (65.0±2%) and *container type* (65.7±6%). As *vessel last region* and *container type* are secondly or thirdly important attributes to form the decision making pathway, it is also important to investigate how these attributes are recorded and associated with the sea container risk profiles.

3.9.8. Decision tree constructions with selected attributes

Table 3-7 shows the reconstructed decision tree classification accuracy using selected attributes (removing, one by one, the less frequently used attributes shown from Section II in Table 3-6). The mean and SD of three test data sets are calculated from each reconstructed decision tree, and are compared with the originally obtained classification accuracy (using all six attributes in Table 3-5) for classification accuracy improvement.

The NZ autumn model shows the highest classification accuracy improvement (66.2±2.5%) using only the first two important attributes from Section II in Table 3-6, *wood type* and *port*, up to 3.5% improvement from the original classification (62.7±2.9%). Interestingly, the classification accuracy drops (63.5±1.6%), when the *port* attribute is removed. This confirms that *wood type* and *port* are significant attributes for the NZ autumn decision tree. The full, NZ spring and NZ winter models show small classification accuracy improvements (< 1% in Table 3-7), which are considered to be insignificant as they may be due to the randomness in the different test data set. However, an interesting point is that the full and the NZ spring models provide similar classification accuracy between using all six attributes and using only two attributes: *wood type* and *container type*. The NZ winter model shows the same classification accuracy between using three attributes (*wood type*, *container type* and *vessel last region*) and using only *wood type* (68.7±2.3), and further removal of the *container type* shows a very small decrease in the classification accuracy (68.4±2.3). There was no scope for improvement for the NZ summer model, as only one attribute, *wood type*, was originally used.

Table 3-7 Classification accuracy (%) of decision tree reconstruction using selected attributes by the TNS method.

Note that the highlighted area indicates where the classification accuracy improvement is found.

Full data		Test sets				
Attributes kept	Attributes removed	1	2	3	Mean	SD
WT, CT, VL, CR, Pol, P.	None (original tree)	64.5	65.8	68.0	66.1	1.8
WT, CT, VL, CR, Pol.	P.	64.5	65.8	68.0	66.1	1.8
WT, CT, VL, CR.	Pol, P.	64.5	65.8	68.2	66.2	1.9
WT, CT, VL.	CR, Pol, P.	65.5	65.8	68.2	66.5	1.5
WT, CT.	VL, CR, Pol, P.	65.5	66.9	68.2	66.9	1.4
WT.	CT, VL, CR, Pol, P.	64.5	66.5	67.6	66.2	1.6
Summer		Test sets				
Attributes kept	Attributes removed	1	2	3	Mean	SD
WT, CT, VL, CR, Pol, P.	None (original tree)	58.0	59.4	62.6	60.2	2.2
WT	CT, VL, CR, Pol, P.	58.5	59.4	62.6	60.2	2.2
Autumn		Test sets				
Attributes kept	Attributes removed	1	2	3	Mean	SD
WT, P, VL, CR, Pol, CT.	None (original tree)	59.3	64.2	64.5	62.7	2.9
WT, P, VL, CR.	Pol, CT.	59.3	64.2	64.5	62.7	2.9
WT, P, VL.	CR, Pol, CT.	62.6	65.9	64.5	64.3	1.7
WT, P.	VL, CR, Pol, CT.	65.0	69.1	64.5	66.2	2.5
WT.	P, VL, CR, Pol, CT.	61.8	65.0	63.7	63.5	1.6
Winter		Test sets				
Attributes kept	Attributes removed	1	2	3	Mean	SD
WT, CT, VL, P, Pol, CR.	None (original tree)	70.9	68.8	66.4	68.7	2.3
WT, CT, VL.	P, Pol, CR.	70.9	68.8	66.4	68.7	2.3
WT, CT.	VL, P, Pol, CR.	70.9	67.9	66.4	68.4	2.3
WT.	CT, VL, P, Pol, CR.	70.9	68.8	66.4	68.7	2.3
Spring		Test sets				
Attributes kept	Attributes removed	1	2	3	Mean	SD
WT, CT, CR, P, Pol, VL.	None (original tree)	64.1	66.1	68.0	66.1	2.0
WT, CT, CR.	P, Pol, VL.	64.1	66.1	68.8	66.3	2.4
WT, CT.	CR, P, Pol, VL.	64.1	69.3	68.8	67.4	2.9
WT.	CT, CR, P, Pol, VL.	63.3	68.5	68.0	66.6	2.9

This experiment shows the need to further investigate attributes to construct good risk profiles. Further classification accuracy improvement is generally observed from using *wood type* and either *container type* or *vessel last region*, but the data set has a seasonal difference, with the NZ autumn data showing the importance of *port* to construct the improved profile.

3.10. Conclusions

Data mining techniques, TNS and TNS-A, were successfully applied as knowledge discovery tools to investigate the decision making process (decision tree) for the sea container contamination profile. A combination of separating out seasonality from data and selected factors by TNS was an alternative approach to extract the maximum knowledge and even improved classification accuracy on the small data set. Results of TNS and TNS-A suggested that each season had unique decision-making processes, although the season may be a proxy for other factors affecting sea container contamination. They extracted hidden knowledge about attributes by detecting how one attribute contributes to contamination in

conjunction with others, to help investigate important attributes and their association with the sea container contamination decision using the generated decision trees.

From here, the future sea container contamination profile and data collection methods may be helped by considering the following: 1) the sea container contamination decision making process shows different seasonal factors and responses, which suggests the need for season-specific profiles; 2) The NZ summer data may need further investigation using different algorithms and attributes, such as social factors, e.g., increasing sea container volume around Christmas; 3) the most important attribute, *wood type*, may require more specific categorisation, e.g., place of origin for the wood or storage place for containers before use; and 4), detailed investigation on the relationship between the *container type*, *vessel last region* and the contamination risk, to understand irregular cases or reasons for misclassification.

Future work could include applying different data mining algorithms, such as fuzzy decision tree techniques, as the result of the decision could have various values, rather than simply being *no* or *yes*. This study was aimed to propose the method that investigated the data freely without losing or modifying the nature of data, but it would be interesting to compare results using binary logic regression even though data points may be reduced further by removing unique instances to produce an applicable training set for validation (Perlich et al. 2003). TNS and TNS-A results may help identify important variables to measure for a cost effective data collection method by discovering hidden knowledge about attributes and the collection of larger data sets over many years in the future would aid in building a more accurate prediction model. An extension from this study, I have made a prototype of a prediction model tool, based on a decision tree algorithm, and introduced the idea to MAF to suggest the creation of an automated early warning system of the sea container contamination risk using the available minimum information about each single container. Since all attributes that are investigated in this study are simple knowledge about the single container that is obtained prior to the ship's arrival in New Zealand, through the documentation of the container form the exporter. It would be an ideal solution, if we can identify high-risk containers even before they enter the country. Details will be discussed in Chapter 7 as future work.

3.11. Acknowledgements

Thanks to Dr. Whyte and the Biosecurity New Zealand Data Analysis team (MAF) for original data collection and processing, advice and support, Dr. H. Cochrane for initiating this project.

3.12. References

Ankerst M, Elsen C, Ester M, Kriegel H-P (2000) Visual classification: an interactive approach to decision tree construction, In Proc. of 7th ACM SIGKDD 01, 392 - 396.

Barlow T, Neville P (2001a) A comparison of 2-D visualizations of hierarchies, In Proc. of INFOVIS 01.

Barlow T, Neville P (2001b) Case Study: Visualization for decision tree analysis in data mining, In Proc. of IEEE symposium on INFOVIS 01, 131 – 138.

Biosecurity Council, Tiakina Aotearoa Protect New Zealand (2003) The Biosecurity Strategy for New Zealand, Biosecurity Council, Wellington.

Border Management Group (2003) Sea container review: MAF Discussion Paper No. 35, Wellington.

Brookmeyer R (2005) Biosecurity and the role of satiations. J R Stat Soc Ser A 168: 263-266.

Freitas AA (2002) Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, Berlin.

ISPM (2002) International Plant Protection Convention, Guidelines for regulating wood packaging material in international trade. International Standard for Phytosanitary Measures, Secretariat of the International Plant Protection Convention, Food and Agriculture Organization of the United Nations Rome, Italy.

Lim TS, Loh WY, Shih YS (2000) A comparison of prediction accuracy, complexity, and training time for thirty-three old and new classification algorithms. Marc Lern 40: 203-228.

Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: a learning-curve analysis. JMLR 4: 211-255.

Ports of Auckland Limited (2006) 50 years inside the box, Auckland, NZ. Available via http://www.poal.co.nz/newsroom/ContainerAnn060421.htm, Accessed on 20 Sep., 2008.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Spate JM, Gibert K, Sànchez-Marrè M, Frank E, Comas J, Athanasiadis I, Letcher R (2006) Data mining as a tool for environmental scientists. In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, Burlington.

Taylor B, Gebbie E, Botherway K, James G, Mormorunni C (2000) New Zealand under Siege: A review of the management of biosecurity risks to the environment, Office of the Parliamentary Commissioner for the Environment, Wellington.

Teoh ST, Ma K-L (2003) PaintingClass: Interactive construction, visualization and exploration of decision trees, ACM SIGKDD 03, Washington DC.

Weiss M, Indurkhya N (1998) Predictive Data Mining: A Practical Guide. Morgan Kaufmann, San Francisco.

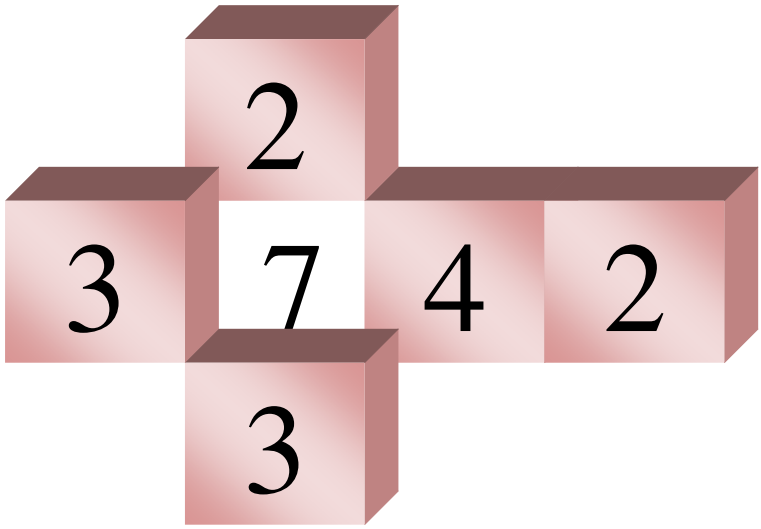
Witten IH, Frank E (2005) Data Mining; Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn, Morgan Kaufmann, San Francisco.

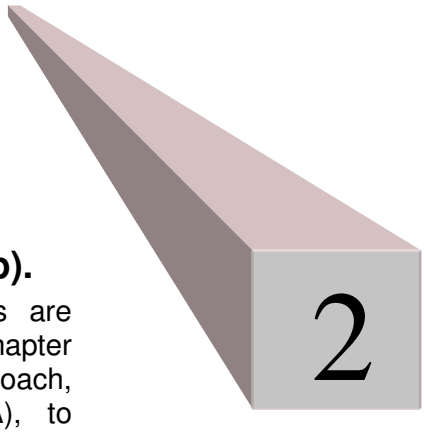
WTO (1995) World Trade Organization, The Agreement on the Application of Sanitary and Phytosanitary Measures, In Results of the Uruguay Round of Multilateral Trade Negotiations: The Legal Texts, Geneva.

3.13. Appendices

Appendix 3-2 Summary of attribute values.

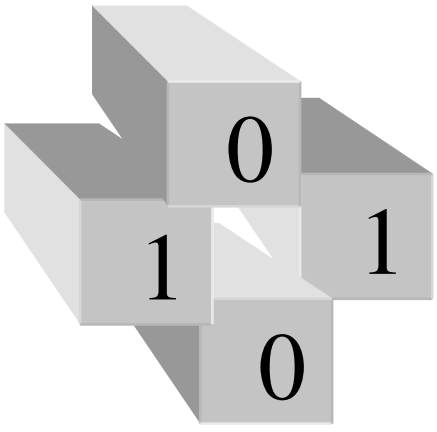
Port (P) Unique variable = 4		Wood type (WT) Unique variable = 4		Container type (CT) variable =8	Unique	Vesse last region (VL) Unique variable = 11		Content region (CR) Unique variable = 12		Contam record	
Auckland	962	Nil	734	General	1234	Australia	740	Australia	418	No	837
Mt Maunganui	195	packaging	398	Reefer	66	Asia Middle East	222	Asia Middle East	282	Yes	564
Lyttelton	195	packaging dunnage	154	Nil	42	South East Asia	198	European Union Scandinavia	175		
Napier	49	dunnage	115	Flat rack	22	North America	118	South East Asia	164		
				Open	20	Pacific Islands	39	North America	110		
				Bulk	9	Japan	22	Nil	102		
				Hazardous	7	Nil	18	Unknown	70		
				Tank	1	Unknown	13	Japan	42		
						EU Scandinavia	12	Africa	20		
						New Zealand	11	Central South America	10		
						Central South America	8	Pacific Islands	7		
								New Zealand	1		
Port of loading (POL) Unique variable = 113											
Melbourne	215	Fremantle	13	Piraeus	4	Tanjung Priok	2	Venezia	1	Madang	1
Singapore	171	Unknown Overseas Port	12	Penang Georgetown	4	Taichung	2	Valparaiso	1	Longbeach	1
Sydney	160	Savannah	12	Jakarta Java	4	Suva	2	Toyama Toyama	1	Livorno	1
Hong Kong	112	Adelaide	11	Bell Bay	4	Sriracha	2	Thessaloniki	1	Lautoka	1
Tanjong Pelepas	64	Houston	10	Yantian	3	Rio Grande	2	Tacoma	1	Kimbe	1
Pusan	57	Tokyo Tokyo	9	Vancouver	3	Port Adelaide	2	Sines	1	Inchon	1
Brisbane	51	Bremerhaven	8	Tuticorin New Tuticorin	3	Norfolk	2	Sao Francisco do Sul	1	Haifa	1
Los Angeles	45	Surabaya Java	7	Toronto	3	Ningbo	2	San Pedro	1	Faaborg	1
Rotterdam	35	Oakland	7	Seattle	3	Nhava Sheva	2	Salerno	1	Dunkerque	1
Port Kelang Port Swettenham	35	Nagoya Aichi	7	Qingdao	3	Nanjing	2	Purfleet	1	Devonport	1
Hamburg	31	Durban	7	Philadelphia	3	Montreal	2	Portland	1	Detroit City	1
Nil	25	Dubai	7	Laem Chabang	3	Mombasa	2	Port Qaboos	1	Columbus	1
Tilbury	21	Gothenburg	6	Kaohsiung	3	Manzanillo	2	Osaka Osaka	1	Cape Town	1
Keelung Chilung	21	Antwerpen	6	Jebel Ali	3	Lisboa	2	Nuku alofa Tongatapu	1	Busum	1
La Spezia	19	Port Kembla	5	Huangpu	3	Karachi	2	Niue Island	1	Burnie	1
Yokohama Kanagawa	16	Pasir Gudang Johor	5	Genoa	3	Izmir Smyrna	2	New Westminster	1	Barcelona	1
Shanghai	16	New York	5	Fos sur Mer	3	Istanbul	2	Muar	1	Ancon	1
Bangkok	14	Le Havre	5	Felixstowe	3	Ho Chi Minh City	2	Memphis	1	Algeciras	1
Kobe Hyogo	13	Southampton	4	Xiamen	2	Wuhan	1	Manila	1		





Chapter 4. Introducing the *K*-Maximum Subarray Algorithm for studying air pollution, climate and health (Fukuda and Takaoka 2007a,b).

Generally, air pollution and health studies are investigated by statistical analyses. This chapter demonstrates the use of a computational approach, the *K*-Maximum Subarray algorithm (*K*-MSA), to identify age cutoff points of acute respiratory admission age groups in relation to current or lagged levels of ambient particulate matter with diameter less than 10 µm (PM_{10}). This method allows exploration of questions like which admission age groups are associated with which PM_{10} levels. This chapter firstly introduces the details of the *K*-MSA concept, and then results will be discussed. The studied data is four years (1998-2002) of daily measurements at neighbourhood scale of admissions (n=1939, 0-98 years of age) and PM_{10} in Christchurch, New Zealand, over varying ages, sexes, annual and winter data with background sulfur dioxide (SO_2) and climate variables. The *K*-MSA detected different dominant and specific admission age cutoff points varying among ages, sexes and season with regard to current or lagged PM_{10}. Identifying such age cutoff points helps defining studied age groups prior to detailed statistical analysis, and increase knowledge about risk assessment to inform the policy making process.



4.1. Introduction

This chapter introduces the impact of air pollution on health problems, briefly covers current air pollution, climate and health research, and demonstrates how the unique computational approach, the *K*-Maximum Subarray Algorithm (*K*-MSA), detects the maximum association of air pollution such as particulate matter (PM), climate and health, measured by the acute respiratory admission rate.

Air pollution has been associated with adverse effects on human health, and increasing mortality and morbidity rates, even when concentrations of ambient air pollutants are below guideline levels (Forsberg et al. 1997; Touloumi et al. 1997; Koening 2000; Anderson et al. 2001; Burnett et al. 2001; Dominici 2002). For example, particulate matter (PM) is made up of small particles, e.g., PM_{10}, with diameter less than 10 µm, whereas a human hair is 50 µm thick. Particulate matter consists of the dust, haze and smoke particles emitted by burning wood, diesel vehicles and industrial operations. This can readily be inhaled and become lodged in the lower lung (ECan 2008), as shown in Fig. 4-1. Even low concentrations of PM, below official guidelines (discussed in Section 5.4. Discussion), can affect human health, e.g., deterioration in pulmonary function, increased respiratory symptoms (chronic cough, bronchitis and chest illness). The combination of sulphur dioxide and particulate matter can increase emergency department visits for asthma (Koening 2000). Thus, estimates of the effects of pollutants on health have now been undertaken in a wide variety of locations, geographies and climates (Wong et al. 1999; Fusco et al. 2001; Erbas and Hyndman 2005; Pope and Dockery 2006). Improving air quality is a priority worldwide, but it is especially important in New Zealand, as an estimated one in four children and one in six adults have asthma (ARF NZ 2007).

4.1.1. Air pollution problem in Christchurch

The study site, Coles Place, is located in a residential area in Christchurch City, in the South Island of New Zealand, adjacent to the Canterbury Plains and Southern Alps, with

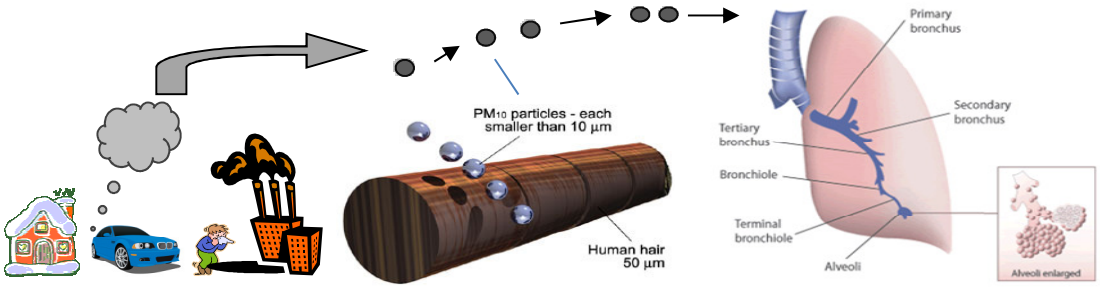


Fig. 4-1 Size of PM_{10} (middle, ECan 2008) and human respiratory system (right, EPA 2008).

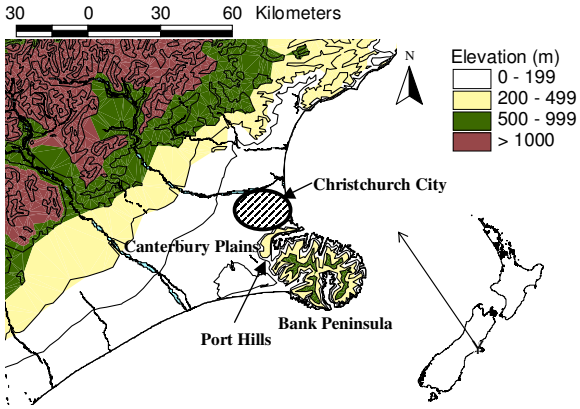


Fig. 4-2 Topography around Christchurch.

small valleys to the west and the Port Hills to the south (Fig. 4-2). The studied area, Christchurch City, has an area of 452 km^2 and a population of about 334,000 people (~123,000 dwellings) (Statistics NZ 2008). The main winter air pollutants in Christchurch are CO_2 from domestic heating and motor vehicles, PM from domestic heating, SO_2 from industry and NO_2 (a product of the oxidation reaction of NO) from motor vehicles (Aberkane *et al*. 2004; Scott and Gunatilake 2004). Additionally, Christchurch researchers Fergusson (1990) and Fergusson and Stewart (1992) stressed the importance of considering effects of the heavy elements, e.g., copper, lead, cadmium, zinc and manganese, on human and environmental health.

The Christchurch local government has proposed to prohibit open fires starting from 2006 (ECan 2008), since Christchurch has a serious winter air pollution problem related to the burning of wood and coal for home heating, combined with poor air dispersion caused by local topographical factors that traps air pollutants in a temperature inversion layer 20-40 m above the ground (Kossmann and Sturman 2004; McKendry et al. 2004). In addition to this, Christchurch winter weather conditions (anticyclonic weather and light northwesterly airflow) cause poor air pollution dispersion (Spronken-Smith et al. 2001). Further, favourable conditions for the accumulation of air pollutants are caused by topographical effects: a zone of flow stagnation, or at least light and variable winds, is caused by the combination of two drainage winds from the Southern Alps and Canterbury Plains, and down the Port Hills of Banks Peninsula (Fig. 4-2) (Kossmann and Sturman 2004).

4.1.2. Air pollution, climate and health research

Estimates of the adverse effects of air pollutants on human health have commonly relied on statistical modelling with time series data. For example, Generalized Additive Models (GAMs) that use nonparametric smoothing techniques to control temporal confounders, e.g., seasonal patterns have been used (Koop and Tole 2006) and Poisson regression to estimate the relationship between fluctuations of daily mortality or morbidity counts and air pollution, while taking into account fluctuations in weather and other time-varying confounders was used by Welty and Zeger (2005). However, the results produced by these models can be highly susceptible to whether researchers have adjusted for potential confounding,

seasonality, weather variables and interactions between pollutants (Lipfert and Wyzga 1995; Samet et al. 2003). Likewise, the choice of regression modelling technique may also markedly alter the estimates between exposure and harm (Erbas and Hyndman 2005). Furthermore, even when researchers use the same technique, there remains considerable scope for variation in study results when varying methods to adjust for confounding are employed (Peng et al. 2006). Previously the studied data was investigated by Fukuda (2004) using a decomposition technique, Singular Spectrum Analysis, which does not require controlling confounding factors by adjusting parameters. It captured detailed oscillation changes such as change points of Christchurch air pollution and climate time series. This decomposition method is suitable for the noisy structures of air pollution and climate time series, to extract the annual, seasonal, daily and hourly local and global climate trends, and their impacts on various air pollution levels as well as their time lag relationship (Fukuda 2004; Fukuda and Hudson 2005). From the later chapter (Chapter 6, Study I) and Fukuda (2007), the decomposition method is found to help the decision tree algorithm predict CO levels using various climate variables.

A different approach is to use Bayesian hierarchical models, which may provide a more complete characterization of the heterogeneity of patient exposure to pollution and weather variables by combining information over various regions and countries (Dominici 2002). This is particularly relevant for these data, because daily winter concentrations of PM_{10} in Christchurch City were found to be non-uniform at the intraurban scale, within a 9.3 km diameter, and epidemiologic studies conducted using the central monitoring site as a proxy for the wider area's exposure may have misclassified daily population exposures (Wilson et al. 2006). Wilson et al. (2006) encouraged more research on outdoor concentration variations at the neighborhood scale (< 4 km) to understand the nature and extent of PM uniformity, considering that a high density of point sources, household chimneys, is one of the sources of winter time pollution in Christchurch. However, environmental or epidemiological data are not always available on a larger scale over many sites. It would therefore be valuable to propose or test a knowledge discovery method that 1) avoids results that are influenced by setting parameter values, and 2) maximizes the utility of small-area collected data to take advantage of the relatively uniform distribution of air pollution in the small area.

4.1.3. The *K*-MSA for air pollution, climate and health study

The the *K*-Maximum Subarray algorithm (*K*-MSA) was developed by Bae and Takaoka (2006) to solve a theoretical computer science problem, improving its performance and parallelizability, and they also explored an image analysis problem (Bae 2007). Fukuda and

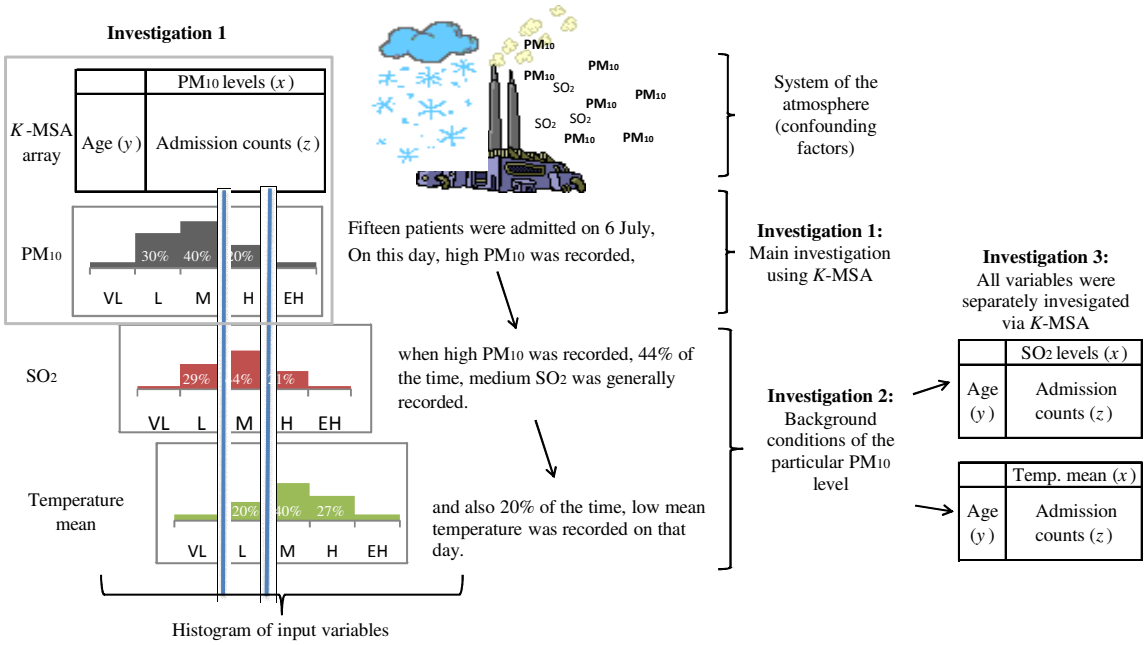


Fig. 4-3 Example demonstration of three investigations in this study.

Takaoka (2007a) used the *K*-MSA for the first time to investigate PM_{10} and the respiratory morbidity rate for summer and winter seasons as a preliminary experiment. Later, Fukuda and Takaoka (2007b) adjusted the *K*-MSA weight parameter to improve the detection of maximum subarray regions to help understanding the maximum relationship between suicide and various social factors.

This study introduces the use of a computational approach, the *K*-MSA, as a knowledge discovery tool, to detect acute respiratory admission age cutoff points in relation to the current or lagged level of ambient particulate matter with diameter less than 10 μm (PM_{10}) by analyzing a 2-dimensional (2-D) array of six PM_{10} classes (divided by the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} percentiles and more) and age groups (five year bands). The *K*-MSA may provide consistent results between investigators and sites, as it is more flexible with sample size and missing values, without specific adjustments, whereas time series analysis generally requires imputing missing data points (Katsouyanni et al. 1996). When the same bin ranges are used to construct the array, then the description of results will be directly comparable among different studies.

In this study, three experiments were carried out; an example demonstration is shown in Fig. 4-3. For investigation 1 (Fig. 4-3), the *K*-MSA inputs were PM_{10} and admission rate, to help understanding the maximum association between the age of the patient and PM_{10} levels. For investigation 2 (Fig. 4-3), seven different potential factors behind the various PM_{10} levels; sulfur dioxide (SO_2), temperature maximum, minimum and average, temperature inversion formation, relative humidity, and rainfall daily measurements, were separately

investigated to indicate the *background conditions* of PM_{10} levels. The system of the atmosphere is complicated; there are many confounding factors that affect the admission rate, and they can be often unknown, indefinable or difficult to determine. Likewise, the regression models, e.g., Poisson regression, directly control or integrate confounding factors as part of the algorithm, whereas the *K*-MSA only takes two input variables (a 2D array) for the analysis. Results from Investigation 1 would show the maximum associations of admissions with patient age groups and PM_{10} ranges. Therefore, to supplement these results, this study investigated the background conditions of other potential variables. For example, fifteen patients were admitted to the hospital on 6 July, and high PM_{10} was detected on this day. High PM_{10} is generally recorded 20% of the time over the studied period. In Investigation 2, histograms of ranges of each potential variable were drawn to investigate the background levels of various climate and air pollution measurements, e.g., SO_2 and temperature mean in Fig. 4-3. When high PM_{10} was generally detected in the study area, Investigation 2 shows 44% of the time, medium SO_2, and 20% of the time, low temperature mean, were recorded. This information helps understanding indirectly how the patient age groups and PM_{10} associate with the various factor levels. Additionally, Investigation 3 (Fig. 4-3) analyzed all seven variables using the *K*-MSA. The following section focuses on results from Investigation 1 and 2 only due to limited space; full results for Investigation 3 are shown in the Appendices. Note that all measurements are strictly taken from the neighborhood scale (< 4 km) so that the distribution of PM_{10} over the study area is approximately uniform.

4.2. Methods

4.2.1. Studied data

Daily measurements of PM_{10}, SO_2 concentrations (in μmg^{-3}), relative humidity (in %), and an indication of the temperature inversion formation (calculated from the difference between the temperatures at 1m and 10m above the ground, with negative values indicating temperature inversion formation) were collected over a four year period (October 1998-September 2002) from a single air pollution monitoring site, located in a medium-size residential area with approximately 12,000 dwellings in northern Christchurch City. Daily measurements of maximum (max), minimum (min) and mean (calculated from the mean of max and min) temperature were taken from another climate monitoring station, located less than 2 km from the air pollution monitoring site.

Over the same period, daily counts of hospital admissions due to respiratory system problems (International Classification of Diseases, ICD-9: 460-519) were obtained for residents domiciled within 2 km of the air pollution monitoring site, aiming to achieve a

Table 4-1 Summary statistics for air pollutants, climate and hospital admissions.

Variables	Mean	SD	Med.	Min.*	Max.
PM_{10} (μgm^{-3})	21.15 (36.69)	21.48 (33.49)	14.8 (24.11)	1.28	207.77
SO_2 (μgm^{-3})	4.71 (8.76)	4.07 (4.58)	3.34 (8.30)	0.00	23.45
T. inversion (°C)	-0.10 (-0.47)	0.76 (0.80)	-0.06 (-0.50)	-3.08	4.60
T. maximum (°C)	17.37 (12.47)	5.18 (3.37)	17.00 (12.00)	3.80	33.90
T. minimum (°C)	7.44 (2.44)	4.68 (3.18)	7.90 (2.30)	-3.80	18.80
T. mean (°C)	12.41 (7.47)	4.52 (2.65)	12.55 (7.25)	1.70	24.55
Relative humidity (%)	73.10 (76.91)	12.33 (11.75)	74.31 (78.08)	31.10	100
Rainfall (mm)	1.65 (1.97)	4.89 (5.53)	0	0	54.00
Hospital admission rate for all age groups					
Female *n*=878 (312)	0.60 (0.85)	0.77 (0.90)	0 (1)	0-1*	4 (0)
Male *n*=1061 (369)	0.73 (1.00)	0.88 (1.03)	1	0-1*	6
Hospital admission rate for age 11-45 years					
Female *n*=148 (47)	0.10 (0.13)	0.32 (0.36)	0	0-0*	2
Male *n*=133 (43)	0.09 (0.12)	0.29 (0.33)	0	0-0*	2
Admitted patient's age range for all age groups					
Female *n*=879 (312)	43.6 (40.7)	32.9 (33.1)	52 (47.5)	4-75*	98 (97)
Male *n*=1061 (369)	42.8 (37.4)	34.3 (34.8)	53 (31)	3-76*	98
Admitted patient's age range for age 11-45 years					
Female *n*=148 (47)	28.8 (27.2)	9.7 (9.3)	28 (27)	20-39 (19.5-33.5)	45
Male *n*=133 (43)	24.3 (22.2)	10.4	23 (18)	14-31 (13-31)	45

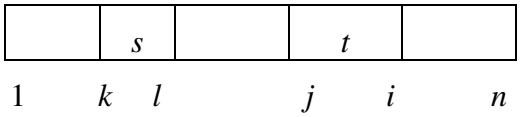
*25th and 75th percentile values are shown for admissions due to minimum values are all zero.
Numbers in brackets indicates winter data. When annual and winter values are the same, one value is shown.

uniform exposure to PM_{10} from point sources for the residents. Daily rainfall measurements (in mm) were only available from 7-8 km away from the air pollution monitoring site, thus are only used as a reference. The studied data contained a maximum of about 4% missing values, mainly from SO_2 and temperature inversion data points. The *K*-MSA does not require a complete data set to form the array, but to make results comparable to past and future investigations, such as time series analysis, imputed data is used (Fukuda 2004).

Summary statistics of air pollution, climate and hospital admissions for the annual (four years) and winter (June to August) data set are shown in Table 4-1. Separate analyses were conducted for females and males, using both the annual and winter data with respect to all age groups (age 0-98 years, *n*=878 for female, and *n*=1061 for male) and just those people aged between 11 to 45 years (*n*=148 for female, *n*=133 for male) to highlight different aspects by excluding the potentially more pollution susceptible young and elderly age groups.

4.2.2. The *K*-Maximum Subarray Analysis

The *K*-Maximum Subarray algorithm (*K*-MSA) written in the C programming language, was developed by Bae and Takaoka (2006, 2007) by enhancing Kadane's algorithm, which finds the maximum subarray of a one-dimensional array, then continuing to develop the 2-D maximum subarray algorithm to locate multiple (*K*) subarrays. Kadane's algorithm, the 2-D maximum subarray problem and the *K*-MSA can be further described thus:



For maximum subarray $a[k..l]$ of $a[1..n]$,
$(k, l) := (0,0)$; $s := -\infty$; $t := 0$; $j := 1$;
for $i := 1$ **to** n **do begin**
$t := t + a[i]$;
  if $t > s$ **then begin** $(k, l) := (j, i)$; $s := t$ **end**;
  if $t < 0$ **then begin** $t := 0$; $j := i + 1$ **end**
end

Fig. 4-4 Diagram to explain Kadane's algorithm (top) and Kadane's algorithm (bottom).

Kadane's algorithm: Let s be the sum of a tentative maximum subarray for the array, $a[k..l]$. Kadane's algorithm has two steps that take $O(n)$ time in total. It scans the given one-dimensional array by accumulating a tentative sum in t. When $t > s$ is detected, s is replaced by t and the position of the maximum subarray so far (k, l) is updated with the position of the tentative maximum subarray (j, i). When $t < 0$ is detected, the accumulation is reset to zero. This process is shown in Fig. 4-4 with the algorithm (Bae and Takaoka 2007).

2-D maximum subarray problem: Let a 2-D array $a[1..m, 1..n]$ be input data, where the value of each element $a[i, j]$ is similar to the one-dimensional problem as seen above. The 2-D maximum subarray problem aims to maximize the sum of the array portion $a[k..i, l..j]$, where (k, l) and (i, j) are index pairs corresponding to the upper-left corner and the bottom right corner of the subarray, described as follows:

1. For each row k of array a ($k \geq 1$)
2. For each row $i \geq k$ of array a
3. Solve the one-dimensional maximum subarray for the strip portion from row k to row i
4. Let the solution be $a[k .. i, l .. j]$
5. Take the maximum of the m(m-1)/2 solutions.

Line 3 takes $O(n)$ time by itself, and it is placed in the doubly nested loop by k and i. Thus, line 3 takes $O(m^2n)$ time in total, and line 5 takes $O(m^2)$ time. Hence, the total time becomes $O(m^2n)$. When $m=n$, this is $O(n^3)$ time, that is, cubic (details in Bae and Takaoka 2006).

The *K* Maximum Subarray problem: The uniqueness of the K-MSA for this application is to identify a cluster of admission counts by incorporating information from forming a 2-D array of respiratory hospital admissions with patient age groups on the vertical (y) index, specific PM_{10} level on the horizontal (x) index, and each matrix cell containing the corresponding count of admissions on the same (or lagged) day; these are generally separately described by 1-D histograms of admission age, time series of admissions and PM_{10} (Fig. 4-5). Each detected maximum subarray describes an admission age cutoff point, i.e.,

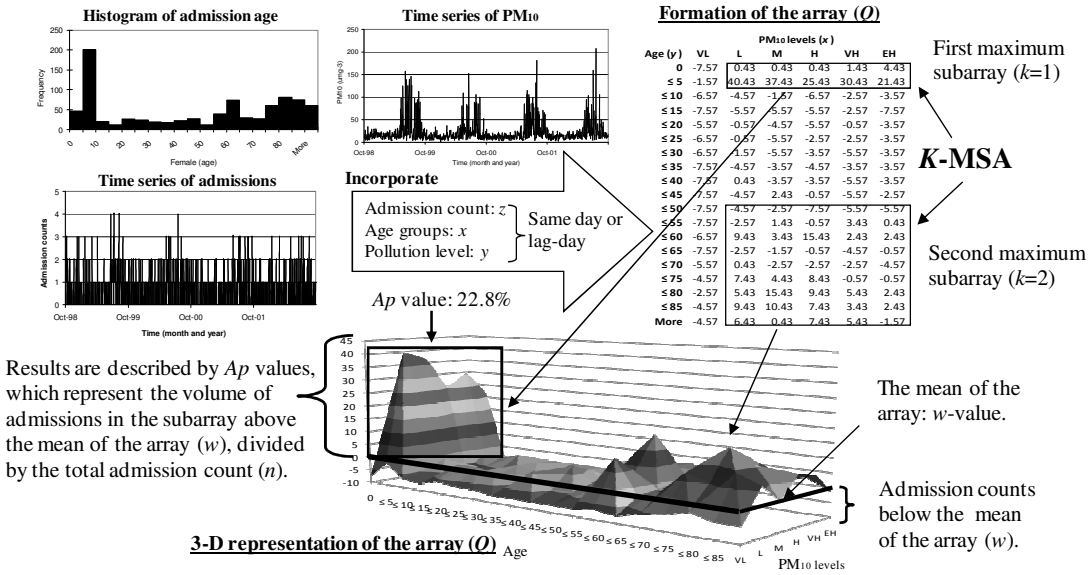


Fig. 4-5 Formation of the array for the *K*-MSA process using female all ages and annual data as an example.

what age groups associate with what levels of PM_{10} and its proportion on the same (or correspondingly lagged) day.

4.2.3. The *K*-MSA for air pollution and health study

Let the horizontal coordinate l be the air pollution range, where $\beta l[1..N\beta l]$ is a series of ranges of air pollution. Let the vertical co-ordinate k represent the patient age range, where $\beta k[1..N\beta k]$ is a series of age ranges. Let $[1, N\beta l]$ be the range in which l lies and $[1, N\beta k]$ be the range in which k lies. In this study, six PM_{10} classes are defined, divided by the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles: *very low* (VL), *low* (L), *medium* (M), *high* (H), *very high* (VH), and *extremely high* (EH), respectively; βl=[VL, L, …, VH, EH] and $N\beta l$=6. An age group is constructed for every five years from 0 to over 85; βk=[0, 5, 10, 15, …, 75, 80, 85, more] and $N\beta k$=19; equivalently, another set of age groups is defined as βk=[15, 20, 25, 30, 35, 40, 45] and $N\beta k$=7, i.e., excluding 0, 5 and 50 years old and more.

A matrix R (dimension $N\beta k$, $N\beta l$) is calculated, such that $R(i, j)$ is the number of observations that lie in the range of (i, j). The rows i of the matrix correspond to age groups, and the columns j correspond to pollution ranges; the cell at the intersection of a row and column contains the number of observations that lie in both of the buckets corresponding to the row and column respectively. Further, the maximum effect of PM_{10} on admission rate at the specific lag is investigated by shifting the admissions data by a Γ-day lag; Γ=[0, 1, 2, .., 6, 7, 14] (note that up to 14 days of air pollution observations are discarded, while the total count of admissions is kept constant). Here, the sum of the array, n, is calculated as

$$n = \sum_{i}^{N\beta k} \sum_{j}^{N\beta l} R(i, j). \tag{4-1}$$

To find a significant maximum subarray, the mean value of R, called the weight parameter (w) is calculated as

$$w = \frac{n}{N\beta k \cdot N\beta l}. \tag{4-2}$$

Finally, a new matrix Q, of the same dimensions as R, is created as,

$$Q(i, j) = R(i, j) - w, \tag{4-3}$$

shown in "Formation of the array (Q)" in Fig. 4-5. The K-MSA detects k^{th} maximum subarrays for k=1, …, K, shown in the box under "Formation of the array" (Fig. 4-5). The speed of the K-MSA is enhanced by extending Kadane's algorithm to be more practical (details in Bae and Takaoka 2006, 2007). When the maximum subarray is detected, all cells in the maximum subarray are replaced with negative infinity (-∞; -X in practice, where X is a large number), and the maximum subarray problem is solved again, on the modified array, which results in $O(Km^2n)$ time, or $O(Kn^3)$ time for m=n. Each subarray is detected without overlapping with other arrays to detect the single array to identify the specific air pollution level and age group.

The concept of the weight value (w) in equation 4-2 was originally developed in Fukuda and Takaoka (2007b) to detect different aspects in detection of maximum subarray. The detailed concept and its purpose will be discussed in Chapter 5, Study I by demonstrating how the detected maximum subarrays with different w-values, e.g., the mean, 75 and 95 percentile, differ from results from the k-means clustering method. Later, Chapter 5, Study II introduces how the use of different w-values detects the generalized positions and centres of the maximum aggregated weed distribution.

4.2.4. Admission proportion in *Ap* values

While the sum of the maximum subarray, s, is the sum of all elements in Q that lie within the maximum subarray, the *proportion of admissions above the w-value*, Ap is calculated as

$$Ap = \frac{s}{n} \times 100\%, \tag{4-4}$$

where n is the total count of admissions over the full array (equation 4-1). The w-value is shown as the thick line in "3-D representation of the array (Q)" in Fig. 4-5. The Ap value represents the proportion of the admissions population that lies above the w-value, which acts as a threshold to highlight the volume of additional admissions above the mean shown in Fig. 4-5. The purpose of this Ap value is to make results comparable between female and male data, which have different sample sizes.

4.3. Results

4.3.1. Maximum effect of PM_{10} with lag

Table 4-2 shows *Ap*, the proportion of admissions above the *w*-value, at *k*=1 for Γ=[0, 1, 2, …, 7, 14]. The mean and standard deviation values (μ ± std) of *Ap* values that were calculated through all Γ values are included as supplementary information. Main discussions will be based on the following rule to select the maximum subarray results at the specific lag (Γ=0). Results are only shown for selected lags due to limited space. The specific lag is selected from observing the greatest *Ap* value (highest admission proportion) at the shortest lag, since the shortest lag may indicate a stronger or more direct impact between the air pollution level and admission rate. If no greater *Ap* value is observed at nonzero lag, the result at Γ=0 is used. If there are several equal greatest values, the shortest lag is used. Selected results (at lag) are shown by highlighted cells in Table 4-2. Note that this selection is based on the first maximum subarray (*k*=1), as it contains the highest admission proportion or majority of the admission rate at the specific lag.

In Table 4-2, all female data show the highest admission proportion (*Ap*=18.9 for annual all age groups, *Ap*=22.8 for winter all age groups, and *Ap*=28.7 for winter age 11-45 years) at the shortest lag, Γ=0, except the annual age 11-45 years data have a significantly high proportion at Γ=5 (*Ap*=26.3 at Γ=5, whereas *Ap*=18.3). For annual all male age groups have the highest proportion (*Ap*=21.6) at the shortest lag, Γ=0. For winter all male age groups have the highest *Ap* at Γ=6 (*Ap*=29.2). For annual and winter male age 11-45 years have the highest *Ap* values at Γ=1 (*Ap*=24.3 and *Ap*=12.6, respectively). The distribution of the admission proportion over different lags for males, in particular all male winter is almost uniform to detect the specific lag at 6-day lag (Γ=6), but annual female age 11-45 years detected the highest Ap value at 5-day lag (Γ=5), when the minimum *Ap* value was detected

Table 4-2 Summary of the admission proportion detected at *k*=1 (%).

Note that all numbers are a proportion of the admission rate (*Ap* in %). The highlighted area is the maximum subarray selected for interpretation (the highest *Ap* value, observed at the shortest lag).

Day of PM_{10}	lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 14	Avg.	SD
All age groups											
Female (annual)	18.9	18.7	18.8	18.9	18.9	18.8	18.3	18.3	18.9	18.7	0.25
(winter)	22.8	20.9	22.4	20.9	20.9	21.2	22.8	20.9	21.0	21.5	0.86
Male (annual)	21.6	21.6	21.5	21.5	21.6	21.1	21.4	20.9	21.5	21.4	0.25
(winter)	28.8	28.8	28.8	28.8	28.8	28.8	29.2	28.8	28.9	28.8	0.13
Age 11-45 years											
Female (annual)	18.3	20.4	18.3	23.6	20.8	26.3	22.3	18.3	19.6	20.9	2.76
(winter)	28.7	25.7	26.6	23.9	20.5	24.8	25.2	26.0	26.0	25.3	2.24
Male (annual)	22.3	24.3	22.2	20.8	20.8	22.3	23.6	22.3	23.1	22.4	1.15
(winter)	10.4	12.6	11.9	9.60	9.60	11.9	11.1	11.1	9.60	10.9	1.15

as 18.3 (at Γ=0, 2 and 7). Thus, the difference in the admission proportion is significantly important for females aged 11-45 years in annual data.

The relatively small range of standard deviation values ($0.13 \leq \text{std} \leq 0.86$) is detected from variations between the sexes, and annual and winter data for the all-ages data, compared with the age 11-45 years data ($1.15 \leq \text{std} \leq 2.76$). The greatest admission proportion (*Ap* value at k=1) is more consistent with the effect of PM_{10} levels on the admission day, but when susceptible age groups are removed (age 11-45), the effect of PM_{10} varies slightly at the specific lag.

4.3.2. Maximum subarray analysis

Results of the first three maximum subarray analyses (up to k=3) at Γ=0 for PM_{10} and selected lags (Section 4.3.1) are summarized for annual and winter data on the left and right side of Table 4-3 (top), respectively. This investigation was described as Investigation 1 in Fig. 4-3. The top part of Table 4-3 shows the *w*-value for each set of data, the maximum subarray number (k), the cumulative sum that is calculated for each *Ap* value up to k=3, and detected age cutoff points or age groups. The horizontal axis indicates the six PM_{10} classes. Highlighted cells in Table 4-3 (top) indicate *Ap* values (representing the proportion of admissions above the *w*-value in %) for each maximum subarray, and the area of the highlighted cell represents the range of PM_{10} levels, from *very low* (VL) to *extremely high* (EH), discussed in following sections, associated with the detected age groups.

The bottom part of Table 4-3 shows background conditions of six SO_2 and climate variables for each PM_{10} class, divided into six levels, from VL to EH, with the same percentile proportions as for PM_{10}. This investigation was described as Investigation 2 in Fig. 4-3. Each number in the bottom part of Table 4-3 shows the frequency (in %) of each class of SO_2 and climate variables with respect to the specific PM_{10} level, and frequencies are rounded to the nearest 1%. For example, the *high* PM_{10} class (H) represents greater than 50^{th} percentile and less than or equal to 75^{th} percentile. In the annual data, when extremely high PM_{10} is observed, the bottom right corner of the left side of Table 4-3 shows that 76% of the time, no rainfall is recorded, 16% of the time, high rainfall is recorded, and so on. Note that details were discussed in Section 4.1.3.

As previously described in details in Section 4.1.3, all *K*-MSA results are described based on the associations between PM_{10} and the admission rate, and the background SO_2 and various climate levels, however, as supplementary information, the maximum subarray results for the individual SO_2 and various climate variables are shown in a similar manner to

Table 4-3 Summary of maximum subarrray results for PM_{10} level and lagged admissions.

Top: Age cutoff points in relation to current-day PM_{10} level (Γ=0) shown in A-J and results at selected lags shown in E-L. Bottom: frequency (in %) of background SO_2 and climate levels against PM_{10} levels.

Annual admissions																							
				Range of PM_{10} ($\mu g m^{-3}$)																			
				≤ 5	≤10	≤ 15	≤ 21	≤ 38	≤ 208														
k	Acc. Sum (%)	Age (years)		Percentile						k	Acc. Sum (%)	Age (years)		Percentile									
				≤ 10th	≤ 25th	≤ 50th	≤ 75th	≤ 95th						≤ 10th	≤ 25th	≤ 50th	≤ 75th	≤ 95th					
				VL	L	M	H	VH	EH					VL	L	M	H	VH	EH				
A. Female lag 0 for all age groups (w=7.58, Γ=0)										G. Female lag 0 for all age groups (w=2.73, Γ=0)													
1		0-5				18.9				1		All				22.8							
2	33.0	≥ 51				14.1				2	24.2	0-5							1.4				
B. Female lag 0 for age 11-45 years (w=3.48, Γ=0)										H. Female lag 0 for age 11-45 years (w=1.12, Γ=0)													
1		16-45				18.3				1		All				28.7							
2	19.3	11-15						1.0		2	32.5	16-20, 31-35			3.8								
										3	34.4	21-25		1.9									
C. Male lag 0 for all age groups (w=9.23, Γ=0)										I. Male lag 0 for all age groups (w=3.24, Γ=0)													
1		0-5				21.6				1		0-5				28.8							
2	39.1	≥ 66				17.5				2	41.6	≥ 66				12.8							
										3	43.1	46-55					1.5						
D. Male lag 0 for age 11-45 years (w=3.17, Γ=0)										J. Male lag 0 for age 11-45 years (w=1.02, Γ=0)													
1		11-35				22.3				1		11-20				10.4							
2	26.6	36-45				4.3				2	12.6	31-40			2.2								
3	28.7	11-15							2.1	3	14.1	26-30					1.5						
											15.6	36-45						1.5					
E. Female lag 5 for age 11-45 years (w=3.48, Γ=5)										K. Male lag 6 for all age groups (w=3.24, Γ=6)													
1		16-45				26.3				1		0-5				29.2							
2	27.4	11-20, 36-40						1.1		2	42.0	≥ 66				12.8							
										3	43.5	11-15				1.5							
F. Male lag 1 for age 11-45 years (w=3.17, Γ=1)										L. Male lag 1 for age 11-45 years (w=1.02, Γ=1)													
1		11-30				24.3				1		11-20				12.6							
2	26.4	11-15							2.1	2	15.6	31-45					3.0						
3	28.4	41-45				2.0				3	17.1	26-40							1.5				
	VL	≤ 0.9		29	13	11	11	4	0		VL	**≤ 2.6**		**24**	**11**	12	3	0	0				
	L	≤ 1.9		22	24	17	16	8	0		L	**≤ 5.3**		**65**	**36**	17	11	3	0				
SO_2	M	≤ 3.3		32	34	28	28	13	1	SO_2	M	≤ 8.1		12	35	27	33	11	7				
($\mu g m^{-3}$)	H	**≤ 6.3**		15	24	33	25	27	7	($\mu g m^{-3}$)	H	≤ 12		0	15	37	35	25	0				
	VH	**≤ 13**		2	6	9	19	**43**	**49**		**VH**	≤ 21		0	4	7	18	**61**	**86**				
	EH	**≤ 23**		0	0	1	1	**5**	**43**		**EH**	≤ 23		0	0	0	0	**0**	**7**				
	VL	**≤ 2**		0	0	0	0	**0**	**0**		**VL**	≤ 5		0	0	1	0	2	0				
	L	**≤ 14**		15	19	14	20	**33**	**67**		**L**	≤ 10		**41**	**21**	23	26	22	21				
Temp.	**M**	**≤ 17**		39	26	24	23	**31**	**27**	Temp.	**M**	≤ 12		**24**	**23**	23	22	27	34				
max.	H	≤ 21		20	28	32	25	21	5	max.	H	≤ 15		0	**27**	23	27	34	28				
(C°)	VH	≤ 32		27	28	30	31	15	1	(C°)	VH	≤ 22		35	**29**	29	23	16	17				
	EH	≤ 34		0	0	0	0	0	0		EH	≤ 34		0	0	1	1	0	0				
	VL	**≤ 2**		2	5	5	10	**25**	**67**		**VL**	≤ -0.05		0	11	12	33	**50**	**52**				
	L	**≤ 4**		5	7	6	10	**18**	**15**		**L**	≤ -0.1		0	0	0	0	0	0				
Temp.	**M**	**≤ 8**		15	25	31	25	**30**	**16**	Temp.	**M**	**≤ 2.3**		6	17	26	25	**30**	**31**				
min.	H	≤ 11		34	36	29	22	15	2	min.	H	**≤ 4.9**		**41**	**32**	29	24	**16**	**10**				
(C°)	VH	≤ 15		41	25	25	27	10	0	(C°)	VH	**≤ 12**		**53**	**40**	33	18	5	7				
	EH	≤ 19		2	3	4	6	2	0		EH	**≤ 19**		**0**	**0**	0	0	0	0				
	VL	**≤ 4**		0	0	0	1	**4**	**19**		**VL**	≤ 3		0	0	1	5	**3**	**3**				
	L	**≤ 9**		12	13	13	19	**33**	**61**		**L**	≤ 6		12	12	15	24	**42**	**41**				
Temp.	**M**	**≤ 13**		20	27	25	22	**31**	**18**	Temp.	**M**	**≤ 7**		24	28	21	24	**27**	**24**				
mean	H	≤ 16		37	33	31	24	19	1	mean	H	**≤ 9**		**29**	**25**	32	23	17	21				
(C°)	VH	≤ 20		29	24	26	28	12	0	(C°)	VH	**≤ 15**		**35**	**35**	29	22	11	10				
	EH	≤ 25		2	4	4	6	1	0		EH	**≤ 25**		**0**	**0**	2	1	0	0				
	VL*	**≤ -0.3**		39	26	29	31	**52**	**82**		**VL***	≤ -0.5		47	29	29	54	**77**	**93**				
	L	≤ -0.6		0	0	0	0	0	0		L	≤ -1.1		0	0	0	0	0	0				
Temp.	M	≤ -0.1		2	13	11	12	14	8	Temp.	M	**≤ -0.5**		0	0	1	1	0	0				
inv.	H	≤ 0.4		32	27	29	29	19	7	inv.	H	**≤ 0.2**		**18**	**32**	23	33	19	7				
(C°)	VH	≤ 2.0		27	32	29	27	14	3	(C°)	VH	**≤ 2.1**		**35**	**39**	47	12	5	0				
	EH	≤ 4.9		0	2	1	1	0	0		EH	**≤ 4.9**		**0**	**0**	0	0	0	0				
	VL	≤ 43		5	3	3	1	0	0		VL	≤ 43		0	0	2	0	0	0				
	L	≤ 66		7	23	27	27	18	20		L	≤ 66		0	18	26	18	14	19				
Relative	M	≤ 74		15	22	28	25	26	27	Relative	M	≤ 74		0	9	16	13	23	29				
humidity	H	≤ 82		20	22	23	24	33	29	humidity	H	**≤ 82**		0	11	16	25	33	30				
(%)	VH	≤ 92		34	25	15	20	23	20	(%)	VH	**≤ 92**		**40**	**43**	26	35	29	16				
	EH	≤ 100		20	6	4	4	0	4		EH	**≤ 100**		**60**	**18**	14	9	1	5				
	No	**0**		29	48	64	70	**75**	**76**		**No**	0		12	28	48	66	**78**	**90**				
Rainfall	H	≤ 0.6		17	11	12	10	11	16	Rainfall	H	**≤ 1.2**		**24**	**21**	24	25	16	7				
(mm)	VH	≤ 1.5		7	7	5	7	5	1	(mm)	VH	**≤ 3**		**6**	**19**	12	1	2	3				
	EH	≤ 54		46	34	19	12	9	6		EH	**≤ 54**		**59**	**32**	16	8	5	0				
Variable	Class	Meas.		Proportions of other variables (%)						Variable	Class	Meas.		Proportions of other variables (%)									

*Strong temperature inversion layer is formed (VL: largest negative value).

PM_{10} results in Table 4-3 top) in Appendix 4-1 for annual data and Appendix 4-2 for winter data. This investigation was described as Investigation 3 in Fig. 4-3.

The first maximum subarray (k=1) detects the greatest *Ap* value and its age cutoff points, called *the dominant admission age groups*, that shows the most susceptible age groups. These dominant admission age groups, generally detected with a broad range of PM_{10}, e.g., *low* to *high*, are perhaps admitted to the hospital regardless of specific changes of PM_{10} levels or due to other factors. The association of the narrower age group and specific PM_{10} levels is detected at a later maximum subarray (k=2 or 3).

4.3.3. Dominant age groups

From the annual current-day PM_{10} effect (Γ=0) shown in Table 4-3, A to D, three dominant admission age groups are detected for both sexes aged 0-5 (female, *Ap*=18.9% at k=1, Table 4-3, A and male, *Ap*=21.6% at k=1, Table 4-3, C), senior age groups of over age 51 years for females (*Ap*=14.1% at k=2, Table 4-3, A) and over age 66 years for males (*Ap*=17.5% at k=2, Table 4-3, C), a wide range of females aged 16-45 (*Ap*=18.3% at k=1, Table 4-3, B) and males aged 11-35 (*Ap*=22.3% at k=1, Table 4-3, D). The larger maximum PM_{10} effect was found from examining age 11-45 years data for females at Γ=5 (Table 4-3, E) and male at Γ=1 (Table 4-3, F). An increase of 8% (*Ap*) for females aged 16-45 was found at Γ=5 (*Ap*=26.3% at k=1, Table 4-3, E) compared with the current-day PM_{10} level (Γ=0). Males did not show such significant differences in proportion and age cutoff at k=1.

The winter current-day PM_{10} effect (Γ=0) on the admission trends shows variations in sexes. Examination of all female age data (0-97 years) shows that all female ages were detected as the dominant admission age group (*Ap*=22.8% at k=1, Table 4-3, G). When the age 11-45 years data were examined, the same all female ages were detected, but the admission proportion increased about 6% (*Ap*=28.7% at k=1, Table 4-3, H) compared with all female age data. This suggests that higher admission proportions are more clustered among younger females than distributed over a wide age range. Male winter data detected identical annual dominant age groups; aged 0-5 (*Ap*=28.8% at k=1, Table 4-3, I) and over age 66 years (*Ap*=12.8% at k=2, Table 4-3, I), but their proportions increased 7.2% and decreased 4.7% respectively compared with annual data. This proportion shift suggested the general dominant admission trend changes; infant and preschool for winter, and senior admission for annual. Additionally, males aged 11-20 is detected (*Ap*=10.4% at k=1, Table 4-3, J). Similarly, male winter data also detected specific lags at 6-day (Table 4-3, K) and 1-day lag (Table 4-3, L), but their age cutoffs and *Ap values* were not significantly different from 0-day lag (Table 4-3, I and J). All age groups that are identified as the dominant age groups with a wide range of

PM_{10} levels may suggest that their associations are due to other factors or with less specific changes in PM_{10} levels.

4.3.4. Annual admission age and specific PM_{10} levels

Generally small admission proportions (*Ap*=1-4%), but specific associations of admission age and annual PM_{10} were detected at later k^{th} subarrays. Variations of age, admission proportion and PM_{10} levels do not show significance between a current-day (Γ= 0) and specific lag. Both sexes aged 11-15 are detected with a high current PM_{10} level (Γ=0); for females, 95th percentile annual PM_{10} (21μmg^{-3} < VH ≤ 38 μmg^{-3}, *Ap*=1.0% at *k*=2, Table 4-3, B), for males, above 95th percentile PM_{10} (EH > 38 μmg^{-3}, *Ap*=2.1 at *k*=3, Table 4-3, D). Additionally, a relatively high admission proportion (*Ap*=4.3%) of males aged 36-45 is detected with 75th percentile annual and current PM_{10} levels (Γ=0, 15 μmg^{-3} < H ≤ 21 μmg^{-3}, *k*=2, Table 4-3, D). Furthermore, females aged 11-20 and 36-40 were detected with 95th percentile PM_{10} at 5-day lag (Γ=5, *Ap*=1.1% at *k*=2, Table 4-3, E) and males aged 41-45 was detected with 50th-75th percentile PM_{10} at 1-day lag (Γ=1, *Ap*=2.0% at *k*=3, Table 4-3, F). However, variations of age and *Ap* values are not significantly different from 0-day lag.

4.3.5. Winter admission age and specific PM_{10} levels

For females, while all female age groups were detected as major admission targets during winter, the association of abnormally high winter current PM_{10} (above 95th percentile (EH) > 95 μmg^{-3}) and females aged 0-5 is detected (Γ=0, *Ap*=1.4%, *k*=2, Table 4-3, G). For male, the association of aged 11-15 and 50th-75th percentile winter PM_{10} levels (15 μmg^{-3} < M, H ≤ 47 μmg^{-3}) is detected at 6-day lag (Γ=6, *Ap*=1.5 at *k*=3, Table 4-3, K). This age group of 11-15 years old for both sexes is also observed previously to associate with the short-term (lag 0-1) of 95th percentile *annual* PM_{10} for female (21 μmg^{-3} < VH ≤ 38 μmg^{-3}) and above 95th percentile *annual* PM_{10} for male (EH > 38 μmg^{-3}).

Detecting such consistent associations of the specific annual and winter PM_{10} levels, e.g., at least greater than 21 μmg^{-3}, for ages 11-15, in particular for males, may associate with specific PM_{10} levels regardless of annual or winter PM_{10} trends. Additionally, a total *Ap* value of 6.7% of four separate male age groups between 26 and 55 years is detected over a range of current PM_{10} levels (Γ=0, 15 μmg^{-3} < M, H, VH ≤ 95 μmg^{-3}, Table 4-3, I and J).

4.3.6. Background SO_2 and climate conditions

The following background SO_2 and climate conditions (bottom in Table 4-3, described in Section 6.3.2) were also noted during observations of the association of above annual and winter 75th percentile PM_{10} levels (includes VH ≤ 95th percentile and EH > 95th percentile) at

0-day lag and specifically both sexes aged 11-15, females aged 0-5, and males aged 36-45 (discussed in Section 6.3.4 and 6.3.5). The total frequencies (in %) of SO_2 levels above 75th percentile (including VH and EH), and climate variables below 50th percentile (including VL, L and M) are summarized when PM_{10} is recorded as VH and EH, respectively. Strong temperature inversion layer formation is observed from large negative values, discussed in Section 6.2.1). Note that all figures for each class that were used to add up to the total frequency are indicated in bold in the bottom part of Table 4-3.

- **SO_2:** when annual SO_2 above 75th percentile (VH > 6.3 μmg^{-3}) is observed, a total of 48% of the time (VH: 43%, EH: 5%), PM_{10} is VH; a total of 92% of the time (VH: 49%, EH: 43%), PM_{10} is EH. When winter SO_2 above 75th percentile (VH > 12 μmg^{-3}) is observed, a total of 61% of the time (VH: 61%, EH: 0%), PM_{10} is VH; a total of 93% of the time (VH: 86%, EH: 7%), PM_{10} is EH.
- **Annual temperature variables:** when daily *maximum* below 50th percentile is observed (M ≤ 17 ºC), a total of 64% of the time (VL: 0%, L: 33%, M: 31%), PM_{10} is VH; a total of 94% of the time (VL: 0%, L: 67%, M: 27%), PM_{10} is EH. When daily *minimum* below 50th percentile is observed (M ≤ 8 ºC), a total of 73% of the time (VL: 25%, L: 18%, M: 30%), PM_{10} is VH; a total of 98% of the time (VL: 67%, L: 15%, M: 16%). When daily *mean* below 50th percentile is observed (M ≤ 13 ºC), a total of 68% of the time (VL: 4%, L: 33%, M: 31%); 98% of the time (VL: 19%, L: 61%, M: 18%), PM_{10} is EH.
- **Winter daily mean temperature:** when daily mean temperature below 50th percentile is observed (M ≤ 7 ºC), a total of 72% of the time (VL: 3%, L: 42%, M: 27%), PM_{10} is VH; a total of 69% of the time (VL: 3%, VL: 41%, M: 24%), PM_{10} is EH.

Note that various levels of temperature maximum are almost evenly distributed over the six PM_{10} classes. Temperature minimum levels separate into two groups, VL and M-VH. When temperature minimum is VL, 50% of the time, PM_{10} is VH, and 52% of the time, PM_{10} is VH. When it is *medium* to *high*, a total of 46% of the time (M: 30%, VH: 16%), PM_{10} is VH, a total of 41% of the time (M: 31%, VH: 10%), PM_{10} is EH, which may cancel out either trend.

- **Strong temperature inversion layer formation (largest negative value at VL**) is observed 52% and 82% of the time when PM_{10} is VH and EH respectively, for *annual* data; 77% and 93% of the time, when PM_{10} is VH and EH respectively, for *winter* data. Higher frequency of temperature inversion is found in winter.
- **Dry conditions (no rainfall**): no rainfall is observed about 75% of the time in annual data, when PM_{10} is both VH and EH; 78% and 90% of the time, PM_{10} is VH and EH in winter, respectively.

Note that relative humidity levels did not show a strong trend; each proportion is almost evenly distributed over six PM_{10} levels during formation of annual and winter PM_{10} VH and EH.

4.3.7. Female admissions and low winter PM_{10}

The unique finding is detected that less than 25 percentile current-day winter PM_{10} ($L \leq 15$ μmg^{-3}) associated with a total *Ap* of 5.7% of admission population for female age 16-25 and 31-35 years (*Ap*=3.8% is the sum of two maximum subarrays detected at *k*=2 from age 16-20 and 31-35 years, and *Ap*=1.9% for age 21-25 years at *k*=3, in Table 4-3, H). This suggests that these female age groups are susceptible to the background conditions when current-day winter PM_{10} is below 25th percentile. Similarly, the total frequency of each SO_2 and climate variable (right side of bottom part in Table 4-3) is summarised below.

- **When winter SO_2 below 25th percentile** is observed ($L \leq 5.3$ μmg^{-3}), a total of 47% of the time (VL: 11%, L: 26%), PM_{10} is L; a total of 89% of the time (VL: 24%, L: 65%), PM_{10} is VL.
- **Winter temperature variables:** when daily minimum above 50th percentile is observed (> 2.3 ºC), a total of 72% of the time (H: 32%, VH: 40%), PM_{10} is L; a total of 94% of the time (H: 41%, VH: 53%), PM_{10} is VL. When daily mean temperature above 50th percentile is observed (> 7 ºC), a total of 60% of the time (H: 25%, VH: 35%), PM_{10} is L; a total of 64% of the time (H: 29%, VH: 35%), PM_{10} is VL.

Note that temperature maximum levels are almost evenly distributed over the six different PM_{10} classes except a total of 65% of the time (L: 41%, M: 24%), PM_{10} is VL.

- **No significant winter temperature inversion layer formation (positive values above 50 percentile at H)** is observed, a total of 71% of the time (H: 32%, VH: 39%), PM_{10} is L; a total of 51% of the time (H: 18%, VH: 35%) of the time, PM_{10} is VL. This indicates that temperature inversion layer is not formed.
- **When winter relative humidity above 75th percentile** is observed (> 82%), a total of 61% of the time (VH: 43, EH: 18), PM_{10} is L; a total of 100% of the time (VH: 40, EH, 60), PM_{10} is VL.
- **When rainfall (above 50th percentile indicates rainfall) is observed,** a total of 72% of the time (H: 21%, VH: 19, VH: 32), PM_{10} is L; a total of 89% of the time (H: 24%, VH: 6, EH: 59%), PM_{10} is VL.

4.4. Discussion

The *K*-MSA was applied as a knowledge discovery tool to detect admission age cutoff points in relation to current PM_{10} levels, and their lag relationships, by forming the maximum admission counts in the array. Since all measurements were taken from the neighborhood scale (< 4 km), this study demonstrates the investigation of an approximately uniform exposure to PM_{10} for residents who were admitted with acute respiratory problems. The *K*-MSA is flexible with respect to sample size and missing values without requiring specific adjustments. However, the *K*-MSA is not an estimation method, and does not directly control for confounding factors. Hence, different background potential confounding factors, e.g., sulfur dioxide (SO_2) and climate variables, were separately investigated with PM_{10}. The objective of the *K*-MSA is to detect *k* maximum subarrays of an array, as opposed to clustering methods, which group points into clusters. Hence, the following chapter (Chapter 5, Study I) will introduce a benchmark experiment using the well known *k*-means clustering method and *K*-MSA to compare results. However, for air pollution and health studies, if the investigator is interested in detecting the *maximum effects* that can be explained by two different factors, the *K*-MSA can be advantageous rather than detecting clusters of points, which do not describe the maximum association. Besides, when the same bin range settings are used to construct the array among different studies, the description of results will be directly comparable. This study is unique, and directly comparable results are not yet available, however, the results from this study generally agree with those obtained using commonly employed analytical techniques such as generalized additive models (Chen et al. 2008).

The *K*-MSA identified that young children and senior age groups were most susceptible to the adverse effects of particulate pollution. Furthermore previous studies have commonly investigated certain pre-defined age groups, for example, persons aged 65 years and older (Schwartz 1996; Medina-Ramón et al. 2006), or divided a population into several age groups: 0-14, 15-64, and ≥ 65, e.g., Anderson et al. (2001). The *K*-MSA can be used as a tool to detect the most appropriate age groups quantified by maximum clustering age cutoff points prior to a statistical analysis. This study detected a cutoff point at 5 years of age and an assessment of the effects of pollution on children of varying age groups seems plausible given the differences in lung function, immune systems and behaviours of children younger or older than 5 years (Moshammer et al. 2006).

Significant effects of PM_{10} on respiratory admission have been detected (Schwartz 1996; Anderson et al. 2001; McGowan et al. 2002), but other researchers have reported that PM_{10} is mainly an indicator of other pollutants (Hagen et al. 2000) and that gaseous air pollutants,

e.g., NO_2, are more important determinants of acute hospitalization for respiratory conditions than particulate mass for Europe (Roemer et al. 1999; Hagen et al. 2000; Fusco et al. 2001). If the array using a different variable detects the sharp association for these dominant admission age groups, then this factor may have a stronger association with admissions compared with PM_{10}.

In this study, admission proportions above the *w*-value of 1-3% (*Ap* value) of a few specific age groups was detected during current formation of high PM_{10} level (lag 0). Firstly, the association of both sexes aged 11-15 and above 75th percentile annual and 50th percentile winter PM_{10} (both levels are equivalent to about > 21 μmg^{-3}) is detected. Similarly, the Australian and New Zealand (including Christchurch) study for children under 14 years old (Barnett et al. 2005) found a positive association of age 5-14 years and PM_{10}, but seasonal differences were not important (RR: 2.1%, 95% CI: -0.2, 4.6 in the cooler and RR: 1.9%, 95% CI: -2.4, 6.4 in the warmer season). Detected age bands are much narrower for our study (11-15 years) than for Barnett et al. (2005), but respiratory admissions of these age groups may associate with particular PM_{10} levels, e.g., > 21 μmg^{-3}, regardless of annual or winter PM_{10}, or warmer or cooler season, although further investigations will be required to conclude this.

Secondly, the association of abnormally high winter PM_{10} (95th percentile; > 95 μmg^{-3}) and female aged 0-5 is detected. Note that the same male age groups are detected as the dominant admission age groups regardless of the specific annual or winter PM_{10}, which may suggest that females aged 0-5 may have a specific response to winter climate conditions behind the formation of abnormally high PM_{10} level, compared with male. Barnett et al. (2005) did not investigate variations among sexes, and investigated only Australian sites for this finding, but they found that when $PM_{2.5}$ and SO_2 were matched with each other, the effect of both pollutants became larger for respiratory admissions for children aged 1-4. Our study shows that these high annual and winter PM_{10} concentrations were observed together relatively frequently; high SO_2 (about > 6 μmg^{-3}; annual 75th and winter 50th percentile), cold temperature (daily annual maximum ≤ 14 ºC, annual min ≤ 2 ºC or winter mean temperature ≤ 7 ºC), strong temperature inversion, and dry conditions (no rainfall for more than 75% of the time). Besides, Christchurch has distinctive pollution sources of PM_{10} and SO_2, domestic home heating and industrial and commercial activities, respectively (Aberkane et al. 2004). Long lasting high air pollutant concentrations, even from different sources of pollutants, can mix and become worse by forming at a low altitude (20-40 m above the ground, McKendry et al. 2004) in a strong temperature inversion layer in winter. Thus, the coupling effects of PM_{10} and SO_2 may have a significant impact on increases in admissions for aged 0-5. Additionally,

Barnett et al. (2005) found that the pollutant impacts on admissions were not related to temperature effects, but their impacts are separate and different. The association of PM_{10} with admissions when matching PM_{10} with temperature (within 1 ºC) was higher than the association of PM_{10} alone with admissions, by approximately 50%, for ages 1-14. However, it is noted that they also found that particulate matter, NO_2 and O_3 pollutants in the warmer season was associated more strongly with respiratory admissions for: age 0 and $PM_{2.5}$; ages 1-4 and both $PM_{2.5}$ and PM_{10}; ages 1-4 and NO_2 and O_3; slightly higher (but not significant) association of ages 1-4 and SO_2 in the cooler season (Barnett et al. 2005).

Interestingly, our study showed a proportion of admissions above the *w*-value of 5.7% (*Ap* value) for females aged 16-25 and 31-35 associated with below 25th percentile winter PM_{10} ($\leq$ 15 μmg^{-3}). This may suggest that these young female adult age groups may be susceptible not to direct winter PM_{10}, but to warmer and wet winter climate conditions; more than 70% of the time, above 50th percentile of daily mean winter temperature (> 7 ºC), high relative humidity (> 78%), and rainfall. Here, the winter pollution problem is the primary focus in Christchurch, but future investigation on warmer season and even warmer days in winter may highlight different aspects.

Some specific male age groups between 26 and 55 years old were found to associate with various high winter PM_{10} levels (15 $\mu mg^{-3} < PM_{10} \leq 95$ μmg^{-3}), but this study consistently noted that females were more susceptible than males; the female senior age cutoff point for annual datasets starts 15 years younger than for males (females over 51; males over 66), and proportions of admissions above the *w*-value of 22-29% (*Ap*) are detected from females of all ages during winter. Similarly, both time series and case-crossover analyses in Ontario show that young (aged 0-14) and adult (aged 15-64) females were more likely to be admitted for air pollution-induced respiratory diseases than males (Luginaah et al. 2005). Also, the frequency of reporting annoyance reactions of air pollutants was higher among people with asthma, women, and people with lack of access to a car, whereas results in this study find that females are generally more susceptible than males (Forsberg et al. 1997).

All the above results and most of our results show the maximum effect of PM_{10} on the respiratory admission rate in the short term (lag 0-1) similar to findings from early studies, e.g., Moolgavkar et al. (1997), Luginaah et al. (2005), Medina-Ramón et al. (2006), although note that the association varies by season, age, sex, and site. Additionally, our study shows an 8% increase in the proportion of admissions above the *w*-value (*Ap*) detected from females aged 16-45, when comparing admissions lagged by 5 days after the PM_{10} level (from 18.3% at 0-day to 26.3% at 5-day lag, but detected PM_{10} levels, a broad range, for both days, did not change). Time series analysis previously conducted on the lag distribution between 0 and 6

for PM_{10} and all respiratory admissions in Christchurch found that slightly higher admissions are detected at lags for each interquartile rise in PM_{10}; 3.37% (95% CI: 2.34, 4.40) with 2-day lag for the highest association, and 3.21% (95% CI: 2.18, 4.24) with 5-day lag for the second highest association, whereas 2.52% with 0-day lag (MacGowan et al. 2002). Some pollutants such as O_3, due to the formation of photochemical smog, are known to have a significant peak in the warmer season (Barnett et al. 2005). A study for the acute effects of O_3 (associated with illness-related absences and especially respiratory absences for an elementary school in Los Angeles) showed an increase at 3-day lag, peaking at 5-day lag, and subsequently a slow decrease (Gilliland et al. 2000). Similarly, a significant increase of respiratory admissions for children under 2 years old in Ontario was detected with a 5-day moving mean of the daily 1-hour maximum O_3 concentration of 45 parts per billion in warmer season (Burnett et al. 2001). The detection of increasing respiratory admission rate at 5-day lag in our study may suggest a possible association with other factors or formation of such peak smog (O_3) at 5-day lag, although our finding was detected for adult females in warm winter days, but ages, sexes and pollutants vary between studies, thus, further investigation will be required to conclude this.

The regional ambient air quality targets for the mean 24 hour period of PM_{10} concentration for good, acceptable (not warranting urgent action), and alert (a warning level) categories are respectively 5-17 μmg^{-3}, 17-33 μmg^{-3}, and greater than 33 μmg^{-3} (Aberkane et al. 2004), whereas this study has detected a specific association of age groups with, for example, greater than 21 μmg^{-3}, 95 μmg^{-3} and even lower than 15 μmg^{-3}. Preferably, reducing air pollutants can improve our health and environmental quality, but recognizing the specific admission trends for the particular age groups, pollution levels and climate conditions, can help provide more accurate advice, e.g., advising limiting or avoiding outdoor activities during smog episodes, targeted to patients by age and sex in order to prevent increased admissions in advance.

4.5. Conclusions

The *K*-MSA was used as a knowledge discovery tool to investigate the acute respiratory admission rate in relation to exposure to PM_{10}, detecting the age cutoff points of its association by forming the maximum clustered admission counts in the array. Proportions of admissions above the *w*-value (*Ap*) of 15-20% were observed for both sexes aged 0-5, females aged 16-45, males aged 11-35, and senior age groups (females over 55; males over 66) regardless of the specific PM_{10} through the year. However, females aged 0-5 associated with abnormally high winter PM_{10} (> 95 μmg^{-3}) and both sexes aged 11-15 associated more

likely with PM_{10} values (> 21 μmg^{-3}) through the year. Females were generally more susceptible than males; in particular admission population above the *w*-value of about 6% (*Ap*) for females aged 16-35 during the formation of low winter PM_{10} ($\leq$ 15 μmg^{-3}), suggested that they may be susceptible more to warmer and wet winter weather conditions (even with low PM_{10}). The most significant maximum association was found at short-term PM_{10} levels (lag 0-1) for both sexes, but a significant increase (about 8%) in the admission population was noted for females aged 16-45 at 5-day lag, suggesting an association with other pollution peaks, e.g., O_3 or NO_2, or other factors.

The *K*-MSA was introduced as a knowledge discovery tool; it is not a statistical method, and does not involve adjusting for confounding factors as do regression models, nor does it assess how results are statistically significant. Hence, the next chapter (Chapter 5, Study II), the spatial weed distribution study, introduces the concept of the randomization test for the *K*-MSA results to assess how the detected maximum subarray regions are statistically or ecologically meaningful. The randomization test is most effective on certain data structures, for example, a large array, with heterogeneous distribution, showing some aggregation, would be likely to have some patterns, usable to quantify the occurrence of randomness among detected maximum subarrays, whereas the small arrays in this case study limit the use of such randomness observations. This will be considered in future. This study can help define age groups prior to detailed statistical analysis, and increase knowledge about risk assessment to inform the policy making process. The *K*-MSA with higher dimensional arrays (3-D to 8-D) will be able to simultaneously assess other pollutants and climate factors.

4.6. Acknowledgements

Thanks to Mr. P. Pearson for processing data and helping with editorial work, Ms.T. Aberkane (ECan) for providing the air pollution and climate data, Dr. A. McDonald and Dr. A. Baumgaertner at the department of Physics, University of Canterbury for providing NIWA climate measurements, and Prof. I. Town, Prof. A. Hornblow, and Assoc. Prof. J. Brown for their advice. The New Zealand Health Information Service originally provided the health data in 2003 for my MSc thesis with grant support from Assoc. Prof. I. Hudson and Dr. C. Williamson; permission was obtained since 2006 for its use in the research at the University of Canterbury by the Canterbury District Health Board (CDHB); URB/06/05/031.

4.7. References

Aberkane T, Harvey M, Webb, M (2004) Annual ambient air quality monitoring report 2002. Report No. U04/57, Environment Canterbury.

Anderson HR, Bremner SA, Atkinson RW, Harrison RM, Walters S (2001) Particulate matter and daily mortality and hospital admissions in the west midlands conurbation of the United Kingdom: associations with fine and coarse particles, black smoke and sulphate. Occup Env Med 58:504-510.

ARF NZ (2007) Learn about asthma. Asthma and Respiratory Foundation of New Zealand (Inc.). Available via www.asthmanz.co.nz. Accessed April 20, 2008.

Bae SE (2007) Sequential and parallel algorithms for the generalized maximum subarray problem. PhD thesis, University of Canterbury, Christchurch.

Bae SE, Takaoka T (2006) Improved algorithms for the *K*-Maximum Subarray problem. Comput J 49:358-374.

Bae SE, Takaoka T (2007) Algorithms for *K*-Disjoint Maximum Subarrays. Int J Found Comput Sci 18:319-339.

Barnett AS, Williams GM, Schwartz J, Neller AH, Best TL, Petroeschevsky A, Simpson, RW (2005) Air pollution and child respiratory health, A Case-Crossover Study in Australia and New Zealand. Am J Respir Crit Care Med 171: 1272-1278.

Burnett RT, Smith-Dorion M, Stieb D, Raizenne ME, Brook JR, Dales RE, Leech JA, Cakmak S, Krewski D (2001) Association between ozone and hospitalization for acute respiratory diseases in children less than 2 years of age. Am J Epidemiol 153:444-452.

Chen Y, Craig L, Krewski D (2008) Air quality risk assessment and management. J Toxicol Environ Health A 71:24-39.

Dominici F (2002) Invited commentary: air pollution and health – what can we learn from a hierarchical approach? Am J Epidemiol 155:11-15.

ECan (2008) *Clean Heat Project*. Environment Canterbury. Available via www.ecan.govt.nz, Accessed April 20, 2008.

EPA (2008) How Does PM Affect Human Health? U.S. Environmental Protection Agency. Available via http://www.epa.gov/region1/airquality/pm-human-health.html. Accessed August 20, 2008.

Erbas B, Hyndman RJ (2005) Sensitivity of the estimated air pollution-respiratory admissions relationship to statistical model choice. Int J Environ Health Res 15: 437-448.

Forsberg B, Stjernberg N, Wall S (1997) People can detect poor air quality well below guideline concentrations: a prevalence study of annoyance reactions and air pollution from traffic. Occup Env Med 54:44-48.

Fergusson JE (1990) The heavy elements: chemistry, environmentl impact and health effects, Pergamon Press, Oxford.

Fergusson JE, Stewart C (1992) The transport of airbone trace elements copper, lead, cadmium, zinc and manganese from a city into rural areas. Sci Tot Environ 121:247-269.

Fukuda K (2004) New improved methods for application and interpretation of singular spectrum analysis: A case study of climate and air pollution in Christchurch, New Zealand. MSc thesis, University of Canterbury, Christchurch.

Fukuda K (2007) Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution. In Proc. of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), 697-704.

Fukuda K, Hudson I (2005) Global and local climatic factors on sulfur dioxide levels: comparison of residential and industrial sites, In Proc. of 20th IWSM 05: 187-194.

Fukuda K, Takaoka T (2007a) Analysis of Air Pollution (PM_{10}) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm, In Proc. of the 2007 ACM SAC 2007, 153-157.

Fukuda K, Takaoka T (2007b) Investigation of the Maximum Association for Suicide Rate and Social Factors Using Computer Algorithm, In Proc. of Modsim 07: 1381-1387.

Fusco D, Forastiere F, Michelozzi P, Spadea T, Ostro B, Arcà M, Perucci CA (2001) Air pollution and hospital admissions for respiratory conditions in Rome, Italy. Eur Respir J 17:1143-1150.

Gilliland FD, Berhane K, Rappaport EB, Thomas DC, Avol E, Gauderman WJ, London SJ, Margolis HG, McConnell R, Islam KT, Peters JM (2000) The effects of ambient air pollution on school absenteeism due to respiratory illness. Epidemiol 12:43-54.

Hagen JA, Nafstad P, Skrondal A, Bjørkly S, Magnus P (2000) Associations between outdoor air pollutants and hospitalization for respiratory diseases. Epidemiol 11:136-140.

Katsouyanni K, Schwartz J, Spix C, Touloumi G, Zmirou D, Zanobetti A, Wojtyniak B, Vonk JM, Tobias A, Pönkä A, Medina S, Bachárová L, Anderson HR (1996) Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. J Epidemiol Community Health 50:S12-18.

Koening JQ (2000) Health effects of ambient air pollution, how safe is the air we breathe? Kluwer Academic, Boston.

Koop G, Tole L (2006) An investigation of thresholds in air pollution – mortality effects. *Environmental Modelling & Software* 21: 1662-1673.

Kossmann M, Sturman AP (2004) The surface wind field during winter smog nights in Christchurch and coastal Canterbury, New Zealand. Int J Climatol 24:93-108.

Lipfert FW, Wyzga RE (1995) Air pollution and mortality: Issues and uncertainties. J Air Waste Manage Assoc 45:949 – 966.

Luginaah IN, Fung KY, Gorey KM, Webster G, Wills C (2005) Association of ambient air pollution with respiratory hospitalization in a government-designated "area of concern": the case of Windsor, Ontario. Environ Health Persp 113:290-296.

McGowan JA, Hider PN, Chacko E, Town GI (2002) Particulate air pollution and hospital admissions in Christchurch, New Zealand. Aust NZ J Public Health 26; 23-29.

McKendry IG, Sturman AP, Vergeiner J (2004) Vertical profiles of particulate matter size distributions during winter domestic burning in Christchurch, New Zealand. Atmos Environ 38:4805-4813.

Medina-Ramón M, Zanobetti A, Schwartz J (2006) The effect of ozone and PM_{10} on hospital admissions for pneumonia and chronic obstructive pulmonary disease: a national multicity study. Am J Epidemiol 163:579-588.

Moolgavkar SH, Luebeck EG, Anderson EL (1997) Air pollution and hospital admissions for respiratory causes in Minneapolis-St.Paul and Birmingham. Epidemiol 8:364-370.

Moshammer H, Bartonova A, Hanke W, van den Hazel P, Koppe JG, Krämer U, Ronchetti R, Spram RJ, Wallis M, Wallner P, Zuurbier M (2006) Air pollution: a threat to the health of our children. Acta Paed Suppl 95:93-105.

Peng DR, Dominici F, Louis TA (2006) Model choice in time series studies of air pollution and mortality. J R Stat Soc Ser A 169:179-203.

Pope CA III, Dockery DW (2006) Health effects of fine particulate air pollution: lines that connect. J Air Waste Manage Assoc 56:709-42.

Roemer W, Clench-Aas J, Englert N, Hoek G, Katsuyanni K, Pekkanen J, Brunekreef B (1999) Inhomogeneity in response to air pollution in European children (PEACE project). Occup Environ Med 56:86-92.

Samet JM, Dominici F, McDermott A, Zeger SL (2003) New problems for an old design: Time series analyses of air pollution and health. Epidemiol 14:11-12.

Schwartz J (1996) Air pollution and hospital admissions for respiratory disease. Epidemiol 7:20-28.

Scott A, Gunatilake M (2004) 2002 Christchurch inventory of emissions to air. R04/03, Environment Canterbury, Christchurch.

Spronken-Smith RA, Sturman AP, Wilton EV (2001) The air pollution problem in Christchurch, New Zealand - Progress and prospects. Clean Air Environ Qual. 36:23-28.

Statistics NZ (2008) Christchurch urban area community profile: at the 2001 Census of Population and Dwelling, Statistics New Zealand. Available via www.stats.govt.nz. Accessed April 20, 2008.

Touloumi G, Katsouyanni K, Zmirou D, Schwartz J, Spix C, Ponce de Leon A, Tobias A, Quennel P, Rabczenko D, Bacharova L, et al. (1997) Short-term effects of ambient oxidant exposure on mortality: A combined analysis within the APHEA project. Am J Epidemiol 146:177-185.

Welty LJ, Zeger SL (2005) Are the acute effects of particulate matter on mortality in the National Morbidity, Mortality, and air pollution study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology* 162: 80-88.

Wilson JG, Kingham S, Sturman AP (2006) Intraurban variations of PM_{10} air pollution in Christchurch, New Zealand: Implications for epidemiological studies. Sci Tot Environ 367:559-572.

Wong TW, Lau TS, Yu TS, Neller A, Wong SL, Tam W, Pang SW (1999) Air pollution and hospital admissions for respiratory and cardiovascular diseases in Hong Kong. Occup Environ Med 56:679 – 683.

4.8. Appendices

Appendix 4-1 The *K*-MSA results for SO_2 and various climate variables, for annual data.

Left: all age groups. Right: From 11 to 45 years old.

All age range							Age (years)	K value	From 11 to 45 years old							Age (years)
	V.L	L	M	H	V.H	Ex.H				V.L	L	M	H	V.H	Ex.H	
A. SO_2: All age groups									**B. SO2: from 11 to 45 years old**							
Female lag 0			22.7%				All	1	Female lag 4				25.6%			16-45
	3.5%						1-5	2						1.0%		11-15
						0.9%	1-5	3			1.0%					36-40
Male lag 0			23.5%				All	1	Male lag 1		24.6%					11-35
	1.8%						1-5	2						2.8%		36-45
		1.5%					≥ 76	3					2.1%			41-45
C. Temp. inversion layer: All age groups									**D. Temp. inversion layer: from 11 to 45 years old**							
Female lag 1	30.0%						All	1	Female lag 6	29.9%						All
			9.9%				0-5	2					7.8%			All
				6.7%			≥ 51	3				3.1%				36-40
Male lag 1	27.6%						All	1	Male lag 1	27.8%						11-40
			11.1%				0-5	2					13.8%			11-35
			9.7%				≥ 61	3					2.0%			41-45
E. Temperature maximum: All age groups									**F. Temperature maximum: from 11 to 45 years old**							
Female lag 4		33.2%					All	1	Female lag 0		35.3%					16-45
								2			1.7%					11-15
								3								
Male lag 0		33.1%					All	1	Male lag 1		33.3%					All
								2								
								3								
G. Temperature minimum: All age groups									**H. Temperature minimum: from 11 to 45 years old**							
Female lag 4	19.1%						0-5	1	Female lag 2			22.9%				16-45
	15.1%						≥ 51	2		2.8%						16-45
			0.2%				21-25	3				0.4%				11-15
Male lag 7	22.2%						0-5	1	Male lag 1				24.1%			11-30
	16.7%						≥ 66	2		4.4%						11-15
			0.8%				51-65	3				2.1%				41-45
I. Temperature average: All age groups									**J. Temperature average: from 11 to 45 years old**							
Female lag 3		27.0%					All	1	Female lag 5		30.5%					16-45
					2.6%		1-5	2				0.4%				11-15
					2.2%		≥ 71	3								
Male lag 7		27.9%					All	1	Male lag 1			**29.8%**				11-35
					3.3%		≥ 66	2			**2.8%**					41-45
					1.8%		1-5	3								
K. Relative humidity: All age groups									**L. Relative humidity: from 11 to 45 years old**							
Female lag 4		28.0%					All	1	Female lag 4		34.0%					16-45
						0.7%	1-5	2			2.4%					11-15
								3								
Male lag 1		27.1%					All	1	Male lag 0		31.3%					11-35
						0.4%	1-5	2			5.5%					36-45
						0.1%	≥ 86	3								
M. Rainfall: All age groups									**N. Rainfall: from 11 to 45 years old**							
Female lag 2	45.9%						All	1	Female lag 2	51.8%						16-45
				6.2%			1-5	2							2.8%	11-20
						3.4%	≥ 71	3					1.4%			21-25
								4							1.4%	31-40
Male lag 3	48.7%						All	1	Male lag 3	60.0%						All
				6.1%			1-5	2							4.3%	11-20
						3.8%	≥ 61	3						0.6%		11-15
	V.L	L	M	H	V.H	Ex.H	Age			V.L	L	M	H	V.H	Ex.H	Age

Appendix 4-2 The *K*-MSA results for SO_2 and various climate variables, for winter data.

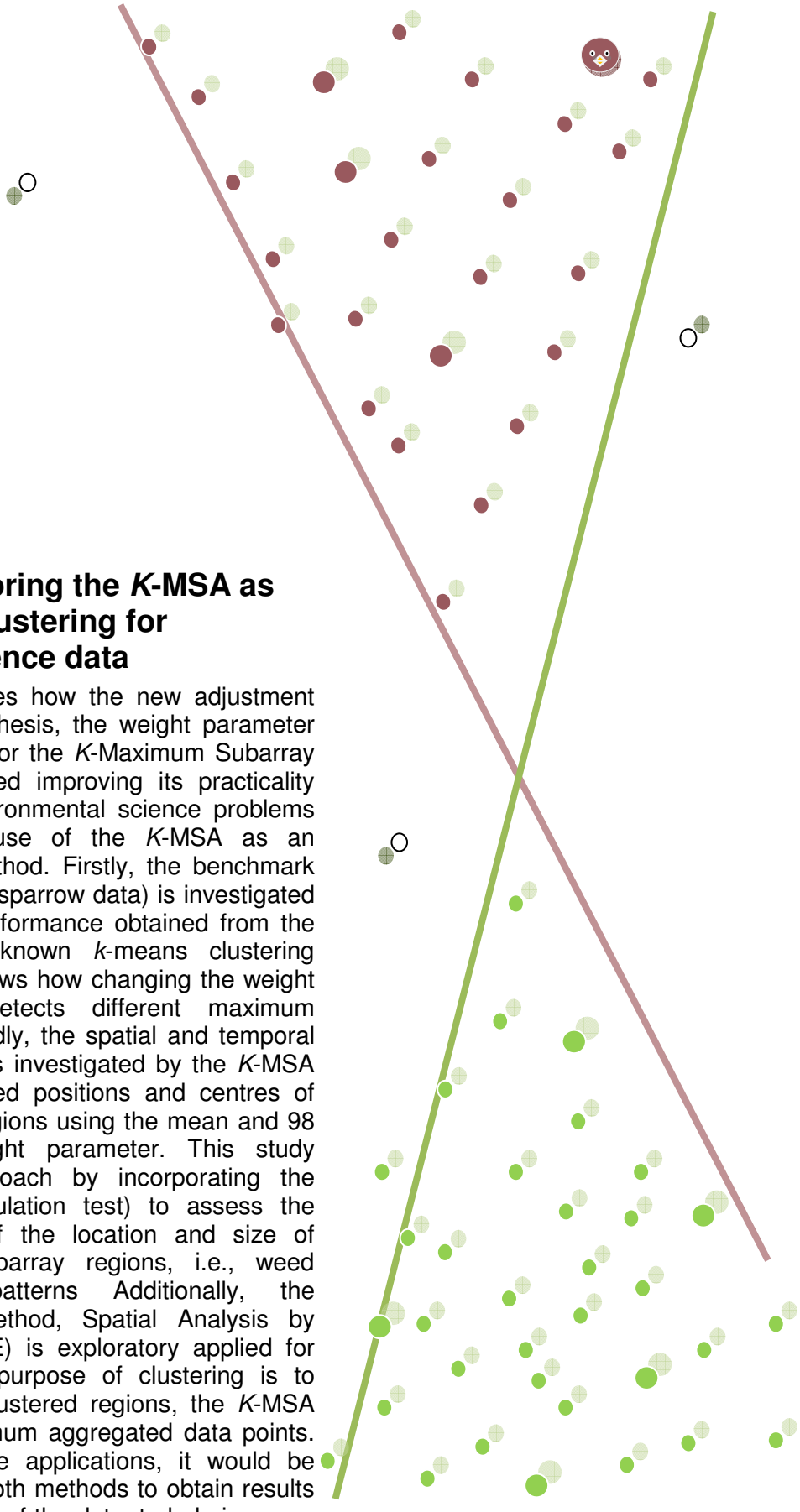
Left: all age groups. Right: From 11 to 45 years old.

All age range	V.L	L	M	H	V.H	Ex.H	Age (years)	K value
A. SO_2: All age groups								
Female lag 6	0.1%		27.2%				All	1
	0.1%						0	2
							56-60	3
Male lag 6			30.7%				0-5	1
			16.1%				≥ 66	2
				1.0%			51-55	3
C. Formation of the inversion layer: All age groups								
Female lag 1#	40.7%						All	1
				10.0%			0-5	2
				4.9%			≥ 51	3
Male lag 5	40.8%						All	1
				14.9%			0-15	2
				4.4%			≥ 71	3
								4
E. Temperature maximum: All age groups								
Female lag 1			32.7%				All	1
								2
								3
Male lag 3			33.3%				All	1
								2
								3
G. Temperature minimum: All age groups								
Female lag 0	9.6%			23.7%			All	1
							All	2
								3
Male lag 0			30.5%				0-5	1
				13.4%			≥66	2
	4.4%						≥71	3
I. Temperature average: All age groups								
Female lag 2			30.8%				All	1
								2
								3
								4
Male lag 2			31.7%				0-5	1
			16.3%				≥ 66	2
			1.0%				11-15	3
K. Relative humidity: All age groups								
Female lag 0*			33.3%				All	1
								2
								3
Male lag 0*			33.3%				All	1
								2
M. Rainfall: All age groups								
Female lag 2		42.9%					All	1
				8.5%			0-5	2
				4.3%			≥ 76	3
Male lag 6		43.5%					All	1
					11.3%		0-5	2
					4.0%		71-80	3
	V.L	L	M	H	V.H	Ex.H	Age	

from 11 to 45 years old	V.L	L	M	H	V.H	Ex.H	Age (years)
B. SO2: from 11 to 45 years old							
Female lag 5				30.2%			11-35
					4.0%		41-45
		1.9%					26-30
Male lag1			11.9%				11-20
			2.9%				26-40
D. Formation of the inversion layer: from 11 to 45 years old							
Female lag 2	38.7%						All
				12.4%			11-35
				1.9%			41-45
Male lag1	11.9%						All
				6.7%			11-20
				0.7%			26-30
					0.7%		31-35
F. Temperature maximum: from 11 to 45 years old							
Female lag 1			36.9%				All
		4.0%					
					4.0%		31-35
Male lag 0			12.6%				11-20
			2.2%				31-40
		0.7%					36-40
H. Temperature minimum: from 11 to 45 years old							
Female lag 0	8.9%			26.7%			16-35
							All
					1.9%		36-40
Male lag 1			11.2%				11-20
				3.0%			36-45
	2.9%						11-35
J. Temperature average: from 11 to 45 years old							
Female lag 1			38.0%				16-45
			1.9%				11-15
					1.9%		31-35
Male lag1				12.7%			11-20
			3.7%				31-45
					1.5%		21-30
L. Relative humidity: from 11 to 45 years old							
Female lag 5		36.9%					All
				9.1%			21-35
					1.9%		11-15
Male lag 7			14.7%				11-40
		0.7%					41-45
N. Rainfall: from 11 to 45 years old							
Female lag 2		51.7%					16-45
				5.6%			21-35
						1.9%	11-15
Male lag 7		19.4%					All
				0.7%			11-15
				0.7%			36-40
	V.L	L	M	H	V.H	Ex.H	Age

* A value of *K* is the same for all lags.

\# A value of *K* at lag 14 is higher (0.413).



Chapter 5. Exploring the *K*-MSA as an alternative to clustering for environmental science data

This chapter introduces how the new adjustment tools developed in this thesis, the weight parameter and randomization test, for the *K*-Maximum Subarray Algorithm (*K*-MSA) helped improving its practicality and explicability for environmental science problems by demonstrating the use of the *K*-MSA as an alternative clustering method. Firstly, the benchmark experiment (the Bumpus sparrow data) is investigated to show the different performance obtained from the *K*-MSA and the well known *k*-means clustering algorithm. This study shows how changing the weight parameter (*w*-value) detects different maximum subarray regions. Secondly, the spatial and temporal weed distribution study is investigated by the *K*-MSA to identify the generalized positions and centres of maximum aggregated regions using the mean and 98 percentile for the weight parameter. This study introduces a new approach by incorporating the randomization test (simulation test) to assess the statistical significance of the location and size of observed maximum subarray regions, i.e., weed spatial aggregation patterns Additionally, the ecological clustering method, Spatial Analysis by Distance IndicEs (SADIE) is exploratory applied for comparison. While the purpose of clustering is to partition the data into clustered regions, the *K*-MSA detects regions of maximum aggregated data points. In environmental science applications, it would be advantageous to apply both methods to obtain results showing different aspects of the data, to help increase knowledge about the data.

Study I. Comparison of the *k*-means clustering algorithm and the *K*-Maximum Subarray algorithm

5.1. Introduction

This chapter covers the brief concept of the *k*-means clustering algorithm, to compare how the *K*-MSA detects maximum subarrays, with how *k*-means builds clusters. Then, a benchmark experiment was carried out on the well-known Bumpus sparrow data using four selected pairs of sparrow measurements based on different underlying structures, examined by multivariate statistics, to demonstrate how the *K*-MSA weight parameter changes the sensitivity of detecting maximum subarrays.

Generally, cluster analysis aims to group or segment a collection of objects into subsets (clusters), such that objects within a cluster are more closely related to one another than to objects assigned to different clusters. This process can be non-hierarchical, when the groups are initially unknown (Johnson and Wichern 2002), but can also arrange the clusters into a natural hierarchy by successively grouping the clusters themselves, thus at each level of the hierarchy, clusters within the same group are more similar to each other than to those in different groups (Hastie et al. 2001).

The *k*-means clustering algorithm is one of the most well-known non-hierarchical clustering methods for detecting clusters and cluster centres in a set of unlabeled or labelled data (Hastie et al. 2001). Over the years, many more clustering techniques have been developed (Tseng and Kao 2006). For example, the pixel classification method (Fukuda and Pearson 2006) that was used for clustering defoliated forest regions in Chapter 4, Study II, is another non-hierarchical method. Popular hierarchical clustering methods are UPGMA (Rohlf 1970), BIRCH (Zhang et al. 1996), CURE (Guha et al. 1998) and ROCK (Guha et al. 1999). Different learning algorithms, such as SOM (Kohonen 1990) for clustering by artificial neural networks, competitive learning networks (Rumelhart and Zipser 1985) and adaptive resonance theory (ART) networks (Carpenter and Grossberg 1987) were developed (details in Tseng and Kao 2006).

In comparison to these clustering methods, the *K*-MSA (Bae and Takaoka 2006; 2007) was not originally developed as a clustering technique, but it acts like an alternative to a clustering method. However, the difference is that the *K*-MSA detects *K* regions (subarrays), in descending order of the sum of the values of the points in each two dimensional array. In order to demonstrate the *K*-MSA as an alternative to clustering, this section compares results of the *k*-means clustering algorithm and the *K*-MSA on the biological benchmark data, the Bumpus sparrow data. Applying the *k*-means clustering and the *K*-MSA algorithm to Bumpus

data is unique, since the detection of clustered groups in the two dimensional array, using in this case two measurements of sparrow morphology, *head and beak* length against *humerus* (bone of the wing) length, identifies how the two factors appear to be clustered or associated. Generally the Bumpus data is used to test various statistical methods, e.g., multivariate statistical analysis or logistic regression (ISU 2008) to prove whether there is significant correlation between factors.

In this study, the Bumpus sparrow data was chosen for the following reasons. Firstly, the underlying structure of the Bumpus sparrow data is well-investigated via various statistical tools (ISU 2008), which helps validating the exploratory investigation. Secondly, the *K*-MSA takes as input a two dimensional array, but is not limited to investigating data that is already in the form of a two dimensional array (e.g., bitmap images); any samples with at least two numerical variables can be processed into a two dimensional array by setting the bin ranges. For example, the previous application in the air pollution, climate and health study (Chapter 4) demonstrated by collecting air pollution data into six bins, labelled from Low to Extremely High and divided by the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile, to detect the association of different air pollution levels (x-axis) and five year bands of respiratory admission age groups (y-axis). When the same percentile ranges are uniformly used by all researchers on all data sets, the detected maximum subarray results should be directly comparable. This investigation also demonstrates another example of applying the *K*-MSA to environmental science data and continuous numerical measurements, as the morphological features of the sparrows are continuous measurements, for example the total body length (mm).

Firstly, this study briefly introduces the history of the Bumpus sparrow data. The Bumpus sparrow data was pre-investigated via multivariate statistics, equicorrelation analysis, Principal Component Analysis (PCA) and Correlation analysis, to validate the underlying data structure. Results from these analyses determined which pairs of sparrow measurements were investigated by *k*-means clustering and the *K*-MSA. Previously, the *K*-MSA was applied with the mean value of the total array used as the weight value, to obtain the generalised maximum subarray to understand the association of air pollution, climate and health (Chapter 6). Here, the concept of the *K*-MSA weight parameter (w-value), originally developed by Fukuda and Takaoka (2007), will be introduced. Setting various weight parameters allows highlighting different aspects of the maximum subarray results to provide extra knowledge about the data. Further, the weed spatial aggregation patterns were investigated with different w-values to identify the generalised positions and centres of maximum aggregated regions, in the next study (Chapter 5, Study II).

5.2. Data and methods

5.2.1. Bumpus sparrow data

English House Sparrows (*Passer domesticus*) were first introduced to the United States in 1850, in New York's Central Park. In January 1898, a number of sparrows affected by a severe storm in Providence, Rhode Island, were brought to the Anatomical Laboratory at Brown University; 72 revived and 64 perished. Hermon Bumpus investigated the reasons for the deaths (Bumpus 1899). Details of the Bumpus data and its further discussions are in Buttermer (1992) and Manly (1994).

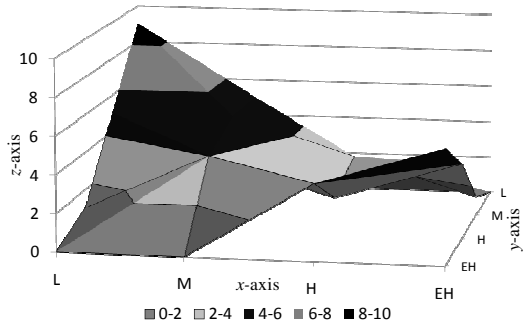
In this study, five distinctive bird morphological features – total length, alar extent (wing spread), length of beak and head, length of humerus (bone of the wing), length of keel of sternum (bone of the middle of the chest) – of the 49 female sparrows (21 surviving and 28 dead) are selected for exploration with *k*-means clustering and the *K*-MSA.

5.2.2. Selection process for four sets of two factors

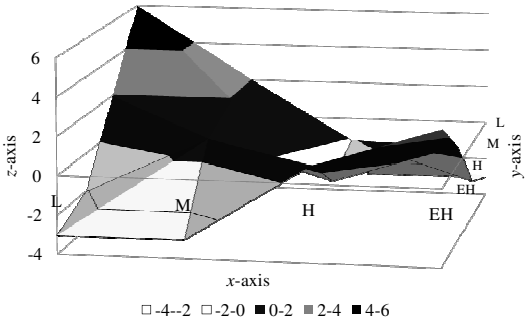
The Bumpus sparrow data were investigated by multivariate statistical analyses, equicorrelation structure test, Principal Component Analysis (PCA) and Correlation analysis, to understand the underlying data structure before testing the *K*-MSA and the *k*-means clustering algorithm. The purpose of this pre-investigation is to help validate and fairly assess the results by investigating the underlying data structure. Results from the statistical analyses are also used to determine which pairs of factors will be used, as for this investigation, the *K*-MSA and *k*-means clustering techniques require the data in two dimensional form. Here, four pairs of factors, with unique underlying structures, are selected: 1) higher PCA component coefficients and correlated, 2) higher PCA component coefficients and less correlated, 3) high and low PCA component coefficients and correlated, and 4) high and low PCA component coefficients and less correlated factors.

Firstly, an equicorrelation structure test was carried out to test whether the data had equal correlation, as the investigation would not work if all factors were equally correlated. Secondly, Principal Component Analysis (PCA) was carried out to identify underlying structures of sparrow measurements, determined by large or small coefficients ($\hat{e}_1$) of factors in the first eigenvector from PCA. A higher coefficient value of a factor, Xi, indicates a higher contribution than a lower coefficient value to the first principal component, where the highest variance (the largest eigenvalue, λ_1) in the data exists. In this data set, the first principal component appears to relate to size differences of sparrows. Furthermore, groups of coefficient values in an eigenvector with the same sign suggest that the corresponding factors share the same influence on the principal component. Lastly, determination of correlation will

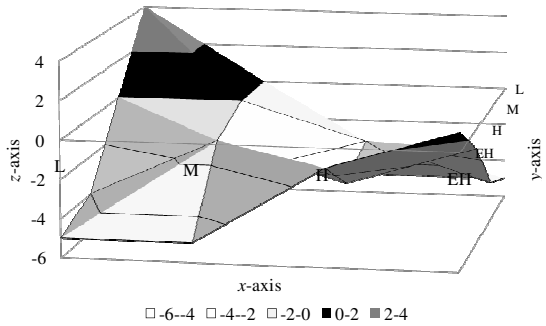
A. Input data



B. Mean weight value (w = 3.06)



C. 75 percentile weight value (w = 5)



D. 95 percentile weight value (w = 9)

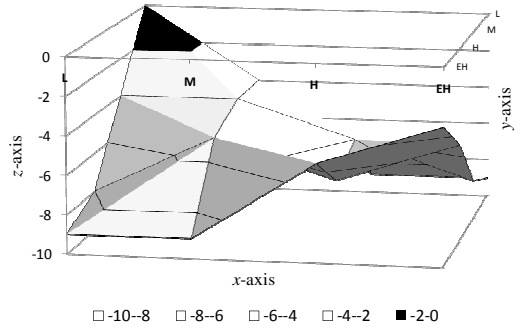


Fig. 5-1 Detected subarray regions with different weight values (w).

Note that solid filled areas (in black) indicated the detected maximum subarrays. The x, y, z-axis is respectively, the length of beak and head (mm), the length of humerus (mm), frequency of two factors categorised in large (L), medium (M), high (H) and extremely high (EH).

identify which pairs of factors are correlated or not. Since the most and least strongly associated or correlated factors may create too strong clustering among features, this study coupled pairs of factors by selecting the second most and least strongly correlated pairs for the methods. Note that the general idea of the statistical analysis is described in Manly (1994).

5.2.3. The *K*-MSA and the weight parameter concept

Details of the *K*-MSA were described in a previous chapter (Chapter 4). The previous application in air pollution and health study (Chapter 4) was demonstrated with the mean of the total array to detect the generalized maximum associations of climate, air pollutants and health. This study further demonstrates how changing the weight value (w) identifies different aspects of maximum subarray results. The concept of changing the weight value was originally developed by Fukuda and Takaoka (2007) to capture specific and more detailed aspects of the maximum subarray for flatter data sets, i.e., without specific or obvious peaks (z-value), to sensitively identify different subarrays over the two-dimensional array. The weight value is like a threshold value; changing it from a lower to higher percentile of the total array allows capturing smaller peaks in the array and allows the experimenter to explore different aspects of maximum subarray results.

From equation 4-2 in Chapter 4, the *K*-MSA detects maximum subarrays that appear above the *w*-value by subtracting the *w*-value from each cell of the total array. Fig. 5-1 demonstrates how the mean, 75 and 95 percentile weight values change the detection regions, and dark coloured areas in Fig. 5-1, B-D indicate the detected maximum subarray regions. Example data were taken with the length of *beak and head* on the *x*-axis, the length of *humerus* on the *y*-axis and the number of sparrows fitting each bin range on the *z*-axis. The *x* and *y* axes are divided into four bins, low (L), medium (M), high (H) and extremely high (EH), divided by the 25^{th}, 50^{th} and 75^{th} percentiles and more than 75^{th} percentiles of each factor, respectively. Basically, each weight value became the ground level of the detection regions, shown at z=0, since the *w*-value was subtracted from each cell, in Fig. 5-1.

The input data is shown in Fig. 5-1, A. Fig. 5-1, B-D shows in darker colours the areas in the data that are above the weight value. Fig. 5-1, B shows that roughly half the data are above the mean. Similarly, Fig. 5-1, C and D illustrate the use of 75 (w = 5) and 95 (w = 9) percentile weight values, showing only the top 25 and 5 percentile of the data respectively. Note that this example is demonstrated by simple structured data. With more complex data structures, i.e., where many more peaks appear, results cannot be easily visualised as in this example. Hence, experimenting with various *w*-values will provide different aspects in the results; how far to explore is up to the experimenter.

In this study, the mean, 75 and 95 percentile of the total array are applied as the weight value, as the input data is reasonably small and has less complicated data structure. However, generally it would be recommended to test different sets of percentiles, e.g., 60, 65, …, 95 percentile, until the experimenter gains feasible results.

5.2.4. The *k*-means clustering algorithm

The *k*-means clustering algorithm is one of the most popular iterative descent clustering methods. Thus, this section briefly describes the algorithm; details are shown in Hastie et al. (2001).

The *k*-means clustering algorithm finds clusters and cluster centres in a set of labelled or unlabeled data by iteratively moving the cluster centres to minimize the total within-cluster variances (unsupervised learning). The former case is applied in this study to detect similar clusters by giving an initial set of centres, and the latter case is used for prediction or classification problems (supervised learning), which is not discussed in this section. Firstly, the *k*-means clustering starts with *K* randomly chosen cluster centres; generally *K* is given by the user. Although, Milligan (1980) stated that *k*-means clustering obtains robust results if the initial cluster centres are provided, since it improves the efficiency of the search process to

avoid spending computation time iteratively testing various k values. Secondly, the algorithm alternately executes the following two steps until convergence (Hastie et al. 2001):

1) *for each data point, the closest cluster centre (in Euclidean distance) is identified;*
2) *each cluster centre is replaced by the coordinatewise average of all data points that are closest to it.*

As with all clustering techniques, the choice of distance or dissimilarity measure between two objects determines clustering.

In this study, the recommendation from Milligan (1980) is taken. While surviving and dead sparrows were known or characterised in this study, two initial centroids are selected randomly from surviving and dead sparrows, respectively. Thus, the initial experiment is carried out to cluster for two physical features of surviving or dead sparrows. In order to avoid biases due to selecting the specific two centroids, the experiment is run more than five times as a check to observe whether the clustering regions would change over five runs by selecting different centroids. All processing in this study is carried out by MINITAB 15 and the standardized variables are applied to make each interpretation as comparable as possible.

5.3. Results and discussion

The aim of this chapter is to compare the k-means and K-MSA algorithms. Thus, results of statistical analyses and detailed meanings of statistical findings will not be discussed, since these are described in Manly (1994). The discussion will be based on the selection of four pairs of factors for the k-means and K-MSA algorithm comparison tests. All statistical analyses were conducted using MINITAB 15.

5.3.1. Selected four pairs of factors for the investigation

Five sparrow measurements, *total length*, *alar extent, beak and head, humerus and keel of sternum* (X_1 to X_5), were investigated as follows. Results of the pre-requisite test, the equicorrelation structure, are shown in detail in Appendix 5-1. This test suggested that total, surviving and dead sparrow data have unequal correlation structures, which confirmed that further statistical analysis was feasible. Each coefficient value of PCA (the first five eigenvectors, coefficients of standardized variables for X_1 to X_5) is shown in Table 5-1. The correlation matrix is shown in Table 5-2. The selection of four pairs of factors using the PCA results was carried out using the first eigenvector from PCA; the first principal component contained about 72.3% (labelled *proportion* in Table 5-1) of the variation in the data, related to differences in sparrow size (Manly 1994).

Table 5-1 Principal component analysis of the total sparrows (*n*=49).

Eigenvalue	3.616	0.532	0.386	0.302	0.165
Proportion	0.723	0.106	0.077	0.060	0.033
Cumulative	0.723	0.829	0.907	0.967	1.000
Variable	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
Total length	0.452	-0.051	-0.690	0.420	0.374
Alar extent	0.462	0.300	-0.341	-0.548	-0.530
Beak and head	0.451	0.325	0.454	0.606	-0.343
Humerus	0.471	0.185	0.411	-0.388	0.652
Keel of sternum	0.398	-0.876	0.178	-0.069	-0.192

Table 5-2 Correlation matrix of total sparrows (*n*=49) using Pearson correlation.

	Total length	Alar extent	Beak and head	Humerus
Alar extent	0.735			
Beak/head	0.662	0.674		
Humerus	0.645	0.769	0.763	
Keel of sternum	0.605	0.529	0.526	0.607

Hence, the largest first coefficients of PCA ($\hat{e}_1$) are observed from *humerus* (0.471), *alar extent* (0.462), *total length* (0.452), *beak and head* (0.451) and slightly lower for *keel of sternum* (0.398), in Table 5-1. The highest correlation was found from *alar extent* and *humerus* (0.769) and the lowest correlation is between *keel of sternum* and *beak and head* (0.526). The secondly correlated and uncorrelated factors, *total length* and *alar extent* (0.735), were selected to form the array containing correlated factors, and *keel of sternum* and *alar extent* (0.529) were selected for uncorrelated factors.

Here, four pairs of factors were selected by combining results from the coefficient values of PCA and correlation analysis:

1) Higher PCA component coefficients and correlated factors; *humerus* and *beak and head*
2) Higher PCA component coefficients and less correlated factors; *humerus* and *total length*
3) Higher and lower PCA component coefficients and correlated factors; *keel of sternum* and *total length*
4) Higher and lower PCA component coefficients and less correlated factors; *keel of sternum* and *alar extent*.

Note that this study does not interpret the sparrow morphology which is examined by the statistical analysis, but details are given in Manly (1994).

5.3.2. Higher PCA component coefficients and correlated factor assessment from the *k*-means clustering and various the *K*-MSA weight parameters

Cluster regions detected for higher PCA component coefficients and correlated factors, *humerus* and *beak and head*, are shown in Fig. 5-2, A for the *k*-means clustering algorithm, and Fig. 5-2, B-D for the *K*-MSA with three different weight values: w = mean, 75 and 95 percentile, respectively. The results of the *K*-MSA are overlaid on the *k*-means clustering plots; label 1 and label 2 for K=2 clustering groups, where one centroid was selected from each surviving (label 1) and dead sparrow (label 2). Each clusters from *k*-means is shown by drawing a line around all its data points (Fig. 5-2, A). The grid lines at the x and y axes indicate the positions of the four bins, Low (L), Medium (M), High (H) and Extremely High (EH), in Fig. 5-2 (25th, 50th and 75th percentiles and more, respectively). Each box in the *K*-MSA plots (Fig. 5-2, B-D) indicates detected subarrays, where all subarrays with a sum above 0, i.e., above the weird value, were used. Note that iteratively selecting two random different centroids over five runs of the *k*-means clustering generally provided the same clustering patterns, so the other results were not shown.

Since the factors in each pair are formed from the higher PCA component coefficients and

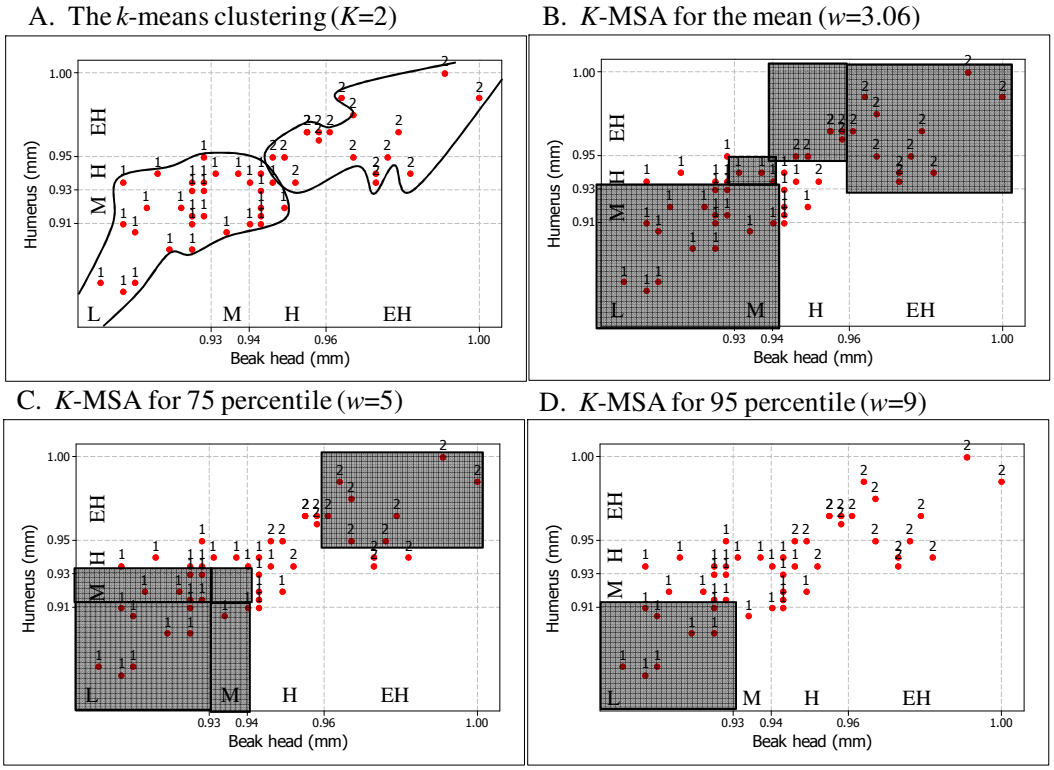


Fig. 5-2 Selected maximum subarray regions using four different weight values (from A to D) for two sparrow measurements, which have higher coefficients and correlated factors.

The *x*-*y* axis is labelled with low (L), medium (M), high (H) and extremely high (EH) measurements, following quartile values.

correlated, the scatter plot shows an almost linear relationship (Fig. 5-2). The *k*-means clusters are detected to separate out two features, the smaller or larger *beak and head* and *humerus* groups, in the middle of the Medium category for both axes. Note that the smaller and larger features of the sparrow groups are detected respectively as surviving sparrows (label 1) and dead sparrows (label 2), though further examination will be required to conclude whether sparrows with the smaller features tended to survive or not. Therefore, this investigation will not discuss the biological meaning of the sparrow features.

The *k*-means clustering clusters all data points into two groups (K=2). However, the *K*-MSA detects maximum clustered regions; since this study only shows results above the *w*-value, not all data points are contained by subarrays. The *k*-means clustering algorithm works by replacing each cluster centre by the coordinatewise average of all data points that are closest to it (Hastie et al. 2001). Clusters from *k*-means clustering (Fig. 5-2, A) are similar to the maximum subarray using the mean weight value (Fig. 5-2, B). The mean weight value *K*-MSA detected four maximum subarrays; two regions at smaller and two regions at larger measurements. An interesting observation is that the border between the two *k*-means clustering groups is detected in almost the same position as the border between the *K*-MSA subarray regions (Fig. 5-2, B). The use of a higher percentile as the weight value for the *K*-MSA (w = 75) split the large maximum subarray covering low *beak and head* and *humerus* measurements into four smaller subarrays (Fig. 5-2, C). Using the 95 percentile weight value detected only one subarray, for smaller measurements (Fig. 5-2, D).

5.3.3. The *k*-means clustering and the mean weight parameter the *K*-MSA

5.3.3.1. Higher and lower PCA component coefficients and less correlated factors

In comparison to the previous case, the scatter plot of higher and lower PCA component coefficients and less correlated factors, *keel of sternum* and *alar extent*, shows large spread (Fig. 5-3), with perhaps no specific trends among these factors. The *k*-means clustering separates such less specific data into two groups by drawing a long diagonal line from the larger *keel of sternum* measurement to the larger *alar extent* measurement to separate the two properly (Fig. 5-3, left). It could be possible that the *k*-means clustering algorithm requires extra clusters ($K > 2$). For example, one data point (circled in Fig. 5-3, left) which is observed at *alar extent* between H and EH and *keel of sternum* less than L, may be an outlier that perhaps should not be grouped with label 1, as the rest of the label 1 group is far from this data point. The *K*-MSA does not detect such a single outlying data point as it only detects maximum clustered subarrays. However, two groups detected from the *k*-means clustering

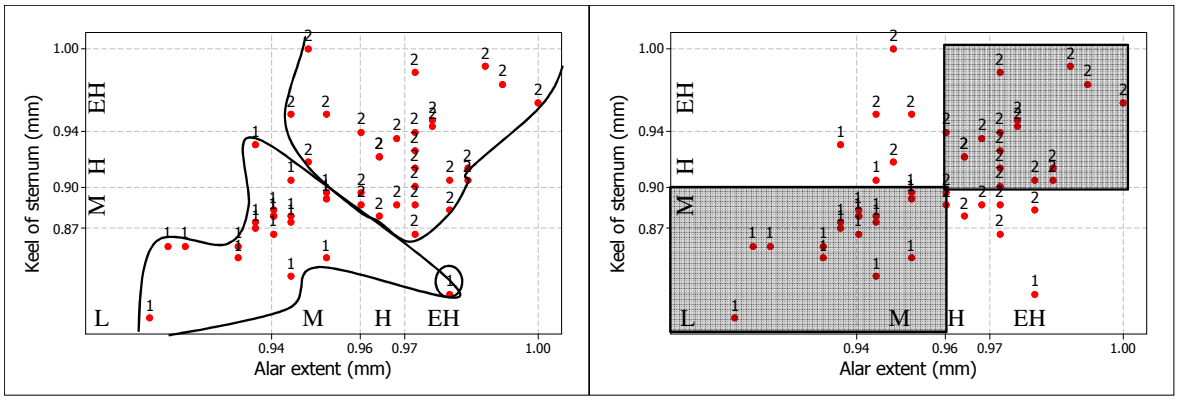


Fig. 5-3 Results of higher and lower coefficients and less correlated factors for the *k*-means clustering algorithm (left) and the *K*-MSA (right).

The *x*-*y* axis is labelled with low (L), medium (M), high (H) and extremely high (EH) measurements following quartile values.

were generally found at similar positions as the *K*-MSA maximum subarrays (Fig. 5-3, right). Overall, it could be said that both methods detect large clustered regions over scattered data points.

5.3.3.2. Higher coefficients and uncorrelated, and higher coefficients and lower coefficients and correlated factors

The scatter plots for higher PCA component coefficients and less correlated factors (*total length* and *humerus*) and higher and lower PCA component coefficients and correlated factors (*total length* and *keel of sternum*) are shown respectively in Fig. 5-4 and Fig. 5-5. Both plots indicate weaker linear trends. The *k*-means clustering and *K*-MSA detected two strongly disjoint clusters, at large and small measurements.

The *k*-means clustering patterns are slightly different between factors with high PCA component coefficients that are less correlated (Fig. 5-4, left) and factors with higher and lower PCA component coefficients that are correlated (Fig. 5-5, left). The former shows that two groups slightly touched at medium (M) level of both factors, and the label 1 cluster is denser while the label 2 cluster is more widely spread. This could possibly be because the two factors have higher PCA component coefficients, but are not quite correlated enough for the cluster groups to be located near each other. Also, two data points with label 2 (circled in Fig. 5-4, left) seem to be better clustered in another cluster – they may require extra clustering groups ($K > 2$), as mentioned earlier. In comparison with this, the *K*-MSA is not forced to cluster all points, thus, a small maximum subarray for label 2 was formed (Fig. 5-4, right), which excluded the potential outliers circled in Fig. 5-4, left. Otherwise, both methods detected a similar large cluster for label 1 (Fig. 5-4).

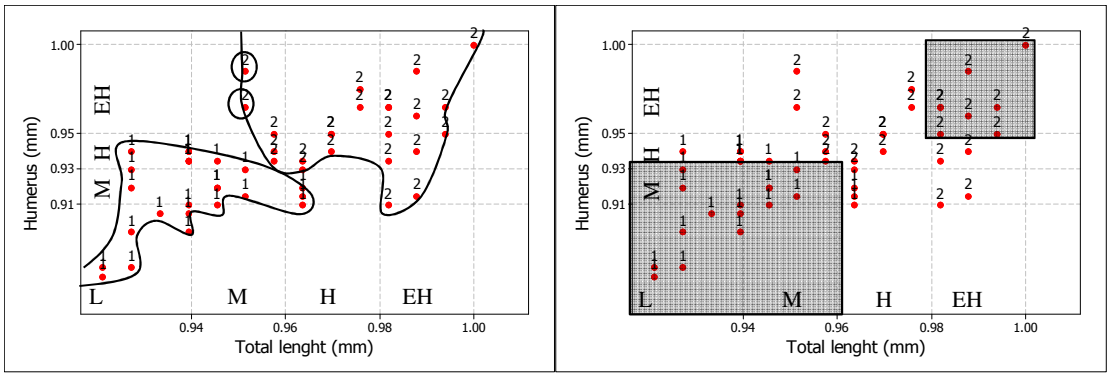


Fig. 5-4 Results of high PCA component coefficients and less correlated factors for the *k*-means clustering algorithm (left) and the *K*-MSA (right).

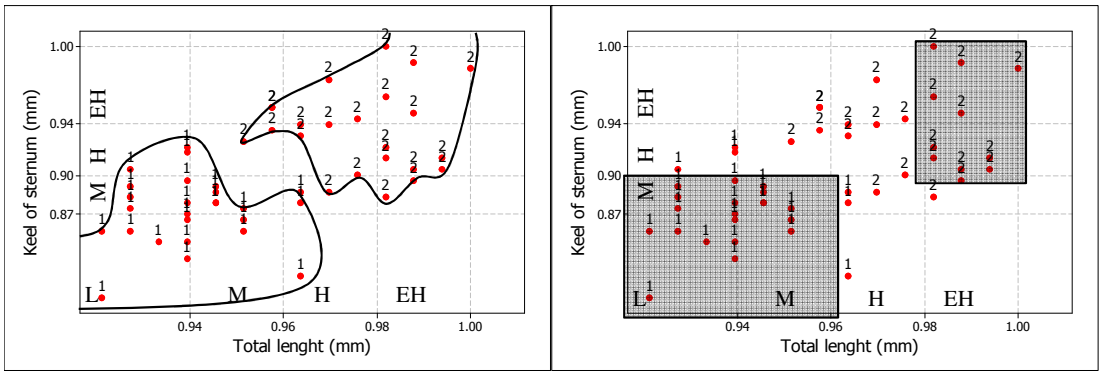


Fig. 5-5 Results of high and low PCA component coefficients and correlated factors for the *k*-means clustering algorithm (right) and the *K*-MSA (right).

The *x*-*y* axis is labelled with low (L), medium (M), high (H) and extremely high (EH) measurements following quartile values.

On the other hand, the latter case, *total length* and *keel of sternum*, shows two separate *k*-means clusters with a distance between them, which could be due to these factors having higher and lower PCA component coefficients, in Fig. 5-5, left. However, these two *k*-means clustered groups seem to show consistent groupings, with two clusters formed at larger or smaller measurements, and their edges seem to capture all data points smoothly (Fig. 5-5, left) compared with the former case (Fig. 5-4, left), maybe because the two factors are correlated. Similar to the former case, both methods detect similar clusters for label 1, but a slightly larger maximum subarray region was detected for label 2 compared with the former case (Fig. 5-5, right). This could be because more data points in label 2 correlated between the two factors; the extremely large *total length* (L) and more than high *keel of sternum* measurements (H and EH), showing more aggregation.

5.4. Conclusions

This study demonstrated, using the benchmark Bumpus sparrow data, the detection of clustering regions via the *k*-means clustering algorithm and maximum subarrays via the *K*-MSA as an alternative to clustering methods. Since *k*-means clustering detects *K* number of

clusters based on the coordinatewise average of all data points that are near them, results of *k*-means clustering and the mean weight value *K*-MSA were found to form similar clustering patterns. The *k*-means clustering algorithm requires a user-provided *K* value (the number of clusters desired), though robust results can be obtained if the initial cluster centres are provided (Milligan 1980). The process will also be more efficient with a known *k* value, to avoid spending computational time iteratively testing various *k* values. While the given *K*-number determines the number of *k*-means clusters, some data points seemed to be forced to group together to form the *K* clusters. As there are many unknown factors involved in environmental science practice, it would be more appropriate to use a method that does not require such *a priori* information as input. Hence, the ecological clustering method, such as Spatial Analysis by Distance IndicEs (SADIE), developed by Perry (1995) and Perry et al. (1999) can be more suitable as it is designed for biological model by considering several ecological data issues; the form of patches, comprising several nearby large counts, and in the form of gaps, comprising several nearby small counts. Whereas SADIE and the *K*-MSA do not require an initial *K* value to start the model, SADIE also searches for clusters based on the average distance between data points, but the *K*-MSA detects different maximum subarrays by selecting various weight parameters, not only detecting the average (maximum) aggregated regions. From here, it will be recommended to test both a clustering method and the *K*-MSA to identify the edge of clustered regions and the maximum aggregated data points, as both methods provide different information to help investigating environmental science data.

From this study, it is suggested that clustering methods, in particular the *k*-means clustering algorithm, detect stable clustering results (the detected edge smoothly contains the clustered data points), when two factors have higher coefficients to results and are correlated. On the other hand, clustering would be more difficult, sometimes forcing points into clusters, if a smaller initial *K* value is given, if the underlying data structure of the two factors has high and lower coefficients to results and the factors are less correlated, as the data is more spread over the two dimensional space. In comparison to the clustering method, the *K*-MSA only detects the regions of maximum aggregated data points regardless of how the data is distributed over the space, thus this may not be a direct issue. The use of higher values of the weight parameter such as 95 percentile would detect only sensitive maximum subarray regions that appear to be significant, such as the top 2 percentile value of aggregated regions.

The next study will demonstrate how the mean and 98 percentile *K*-MSA weight values are applied to detect the spatial and temporal weed aggregation pattern. Additionally, SADIE will be applied to demonstrate different performance in detecting the weed clustering regions.

5.5. References

Bae SE, Takaoka T (2006) Improved algorithms for the *K*-Maximum Subarray problem. Comput J 49:358-374.

Bae SE, Takaoka T (2007) Algorithms for *K*-Disjoint Maximum Subarrays. Int J Found Comput Sci 18:319-339.

Bumpus HC (1899) The elimination of the unfit as illustrated by the introduced House Sparrow, Passer domesticus, Biol. Lectures, Marine Biol. Lab., Woods Hole: 209-226.

Buttermer WA (1992) Differential overnight survival by Bumpus' house sparrows: an alternate interpretation, The Condor 94:944-954.

Carpenter GA, Grossberg S (1987) ART 2: self-organization of stable category recognition codes for analog input patterns, Appl Opt. 26: 4919-4930.

Fukuda K, Pearson PA (2006) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-6.

Fukuda K, Takaoka T (2007) Investigation of the maximum association for suicide rate and social factors using Computer Algorithm, In Oxley L and Kulasiri D. (eds) MODSIM 07, 1381-1387.

Guha *S*, Rastogi R, Shim K (1998) CURE: An efficient clustering algorithm for large databases. In Prof. of the 1998 ACM SIGMOD Intl Conf on Management of Data: 73-84.

Guha *S*, Rastogi R, Shim K (1999) ROCK: A robust clustering algorithm for categorical attributes. In Proc. of the 15th Intl Conf. on Data Engineering: 512-521.

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning, data mining, inference, and prediction. Springer, Canada.

ISU (2008) Output from SAS Logistic Regression Analysis of Bumpus' Data. Iowa State University of Science and Technology. Available via http://www.public.iastate.edu/~fjanzen/software/regress.htm#output, Accessed on August 17, 2008.

Johnson R, Wichern D (2002) Applied multivariate statistical methods, 5th edt. Prentice Hall, New Jersey.

Kohonen T (1990) The self-organizing map. In Proc. of the IEEE 78:1464-1479.

Manly BFJ (1994) Multivariate statistical methods, A primer. Chapman & Hall, London, 2nd ed.

Milligan GW (1980) An examination of the effect of six types of error pertubation on fifteen clustering algorithms. Psychom 45: 325-342.

Perry JN (1995) Spatial Analysis by Distance Indices. J Anim Ecol 64; 303-314.

Perry JN, Winder L, Holland JM, Alston RD (1999) Red-blue plots for detecting clusters in count data. Ecol Let 2 :106-113.

Rohlf FJ (1970) Adaptive hierarchical clustering schemes. Syst Zool 19:58-82.

Rumelhart DE, Zipser D (1985) Feature discovery by competitive learning, Cogn Sci 9:75-112.

Tseng VS, Kao C-P (2006) Chapter IX Parameterless clustering techniques for gene expression analysis. In Hsu (eds) Advanced data mining technologies in biorinformatics. IDEA Group Publishing, Hershey.

Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for very large databases. In Prof. of the 1998 ACM SIGMOD Intl Conf on Management of Data: 103-114.

5.6. Appendices

Appendix 5-1 Analysis of equicorrelation for the Bumpus sparrow data.

The equicorrelation structure, followed by Johnson and Wichern (2002), determined the equicorrelation structure for total, surviving and dead sparrows with 5% critical value using a chi-square distribution; *df*=1/2 (*p*+1)(*p*-1). This study examined up to five eigenvalues, *p*=5, thus the 9% critical value for this hypothesis test is χ_9^2 (α=0.05) = 8.34.

The hypothesis (equal correlation: ρ_0) is rejected (equation 5-1), when the *T*-value is larger than the chi-square distribution (equation 5-2).

$$H_0: \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} \qquad \textbf{(5-1)}$$

$$H_1: \rho \neq \rho_0$$

$$T = \frac{(n-1)}{(1-\bar{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^{-2} - \hat{\gamma}\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2\right] > \chi_{(p+1)(p-2)/2}{}^2(\alpha) \qquad (5\text{-}2)$$

Results in Table 5-3 confirmed that total, surviving and dead sparrow data have unequal correlation structures, since all tests have *T*-values larger than the critical value (*T*=15.24 for total, *T*=10.28 for surviving and *T*=10.45 for dead). Note that the sample correlation matrix, *R*, for equation 5-2 is shown in Table 5-3.

Table 5-3 Determination of equicorrelation structure for total, survived and dead sparrows with 5% critical value using a chi-square distribution.

Total sparrows (*n*=49): *T*= 15.2414	Surviving sparrows (*n*=21): *T*=10.2799	Dead sparrows (*n*=28): *T*=10.4994
R =	R =	R =
1.0000 0.7350 0.6618 0.6453 0.6051	1.0000 0.6545 0.6425 0.6239 0.5104	1.0000 0.7762 0.6770 0.6824 0.6569
0.7350 1.0000 0.6737 0.7685 0.5290	0.6545 1.0000 0.6264 0.7464 0.2774	0.7762 1.0000 0.6978 0.7846 0.6200
0.6618 0.6737 1.0000 0.7632 0.5263	0.6425 0.6264 1.0000 0.6180 0.4336	0.6770 0.6978 1.0000 0.8347 0.5699
0.6453 0.7685 0.7632 1.0000 0.6066	0.6239 0.7464 0.6180 1.0000 0.4165	0.6824 0.7846 0.8347 1.0000 0.6678
0.6051 0.5290 0.5263 0.6066 1.0000	0.5104 0.2774 0.4336 0.4165 1.0000	0.6569 0.6200 0.5699 0.6678 1.0000

Study II. Investigation of spatial weed distribution using the *K*-Maximum subarray

5.7. Introduction

This chapter covers the brief concept of the weed management problem, introduces the application of the *K*-MSA for spatial analysis, discusses in detail the effective use of different weight parameters to highlight different aspects in detecting weed aggregation patterns and locations, the idea of incorporating statistical analysis, the randomisation test, and the ecological clustering method, Spatial Analysis by Distance IndicEs (SADIE, Perry 1995; Perry et al. 1999), for comparison with *K*-MSA results.

Weed management and control aims to minimize weed impact on the environment and human activities, including agricultural and recreational activities. However, maintaining the health of the surrounding ecosystem without altering its nature can be challenging. The spread of herbicides can cause unwanted damage to other crops and the ecosystem, and herbicides may be an expensive option. Some weeds, e.g., *Striga hermonthica*, commonly known as witch weed, cause most of their damage while underground, with seeds that remain dormant in the soil. These weeds are hard to control by direct herbicide application (Hess et al. 2001). Effective weed management often involves long term control strategies. Identification of the driving factors of weed spread is an important goal, to help mitigate and control further weed invasion. The role of changing drivers of plant invasions has received little attention at the intra-specific level, and the factors determining weed spread are not fully understood (Dietz and Edwards 2006; Hess et al. 2001; Williams et al. 2007).

Common practices of describing weed spatial distribution vary, for example, from lab experiments to understand the theoretical mechanism of wind for weed seed dispersal, e.g., Wang et al. (2008) and Hess et al. (2001), to field based studies that integrate hydrological and landscape information, e.g., Swetnam et al. (1998).

In this study, the data were available for the naturalized tree, hawthorn (*Crataegus monogyna*) in Porters Pass, in central Canterbury, New Zealand. Hawthorn was in the past impeded by grazing, but has more recently spread and is considered to be a serious environmental threat. Previously, Williams and Buxton (1986) investigated the dynamics of hawthorn by determining biological indexes such as age structure, growth rate, stem diameter and height. To date, spatial quantitative analysis on these data has not been undertaken (Williams, personal communication, 4 Aug, 2008). Hence, a purpose of this study was to introduce the unique computer algorithm, the *K*-Maximum Subarray Algorithm (*K*-MSA) as

a clustering method to help quantifying and understanding the spatial distribution of hawthorn.

5.7.1. Motivations of the *K*-MSA application for the spatial ecological data

The previous benchmark experiment suggested that the *K*-MSA produces similar clustering points to the widely used *k*-means clustering (in Chapter 5: Study I). A possible advantage of using the *K*-MSA (Bae and Takaoka 2006; 2007) for ecological studies is that the *K*-MSA detects unlabelled *K*-number of *maximum* clustered regions, corresponding to maximum aggregations of hawthorn, whereas the *k*-means clustering techniques detect the edge of clustered regions. The application of the *K*-MSA to hawthorn provides information about temporal and spatial changes of maximum aggregated regions, detecting how the population and position of the most highly aggregated regions have been changing. Applying the concept of changing the weight parameter to the mean and 98 percentile values of the total array (Fukuda and Takaoka 2007) allows investigating the generalised positions and centres of maximum aggregated hawthorn regions.

In this study, four different periods of hawthorn distribution patterns were investigated: 1966, 1976, 1986 and 2006. Three separate examinations were carried out to demonstrate how the *K*-MSA can help future weed management strategy by quantifying its spatial distribution for ecological data. Firstly, the *K*-MSA was used to investigate the spatial hawthorn distribution by detecting the position of the maximum aggregated hawthorn regions using different weight values. Secondly, a randomisation test was carried out to calculate the statistical significance of the cluster regions detected by the *K*-MSA. Thirdly, the *K*-MSA results were compared with results for Spatial Analysis by Distance IndicEs (SADIE), developed by Perry (1995) and Perry et al. (1999). SADIE is a similar technique to the *K*-MSA and was developed as a spatial analysis tool for ecological data, to detect clusters in the form of patches, comprising several nearby large counts, and in the form of gaps, comprising several nearby small counts.

The uniqueness of this studied data set is that historical data is available since 1966. Results from 1966 and the known position of the hawthorn origin can act like ground truth. Thus, detected results from the later periods can be validated as to how the maximum regions were changing or moving relative to the origin. If the maximum aggregated region is assumed to be at or near the origin (no weed management or control has been carried out on the site), the *K*-MSA can be a tool to help identifying such a mechanism for historically unknown data. However, it is more likely feasible to suggest that the *K*-MSA can be a tool to

detect the maximum aggregated regions to help increasing knowledge about the hawthorn distribution pattern to help the weed management strategy.

5.8. Methods

5.8.1. Studied data

Hawthorn is a European shrub or low tree with white flowers. It was first recorded in the wild in New Zealand in 1899 and widely distributed in the South Island, though some regional councils of New Zealand have classified it as a noxious plant (Williams and Buxton 1986). Hawthorn is known to reach reproductive age slowly, but produces abundant fleshy fruit. Seed dispersal is primarily by European blackbirds, and seedlings are only partially grazing resistant (Williams et al. 2007). The data set was collected and provided by Dr. Williams and his colleagues at Landcare Research and AgResearch. Specific survey methods are beyond this study; the detailed survey method for this dataset was published in Williams and Buxton (1986). A brief history of the data set and location is introduced from Williams and Buxton (1986) and Williams et al. (2007).

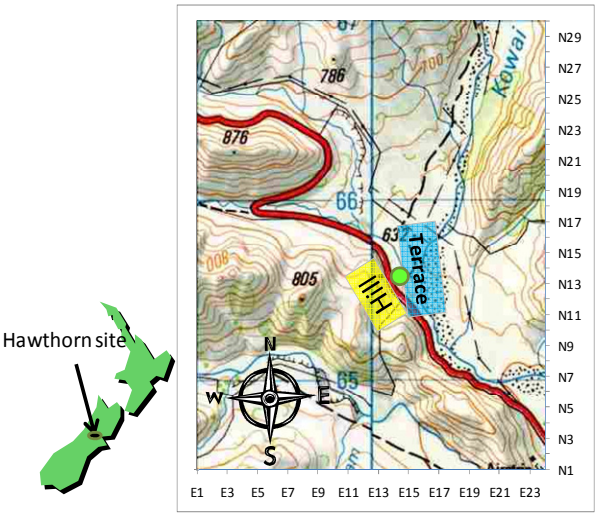


Fig. 5-6 Hawthorn study site and the origin of the hawthorn in 1930 (green dot), hill site (yellow) and terrace site (blue).

Note that the *x*- and *y*-axis indicate easting and northing respectively.

Table 5-4 Mean and 98 percentile values of each hawthorn data set for the weight parameter.

Studed period (year)	1966	1976	1986	2006
Number of hawthorn trees (n)	32	192	320	523
w = mean	0.04	0.27	0.45	0.73
w = 98 percentile	1	2	4	7

Williams et al. (2007) mentioned that hawthorn was first recorded in the wild in New Zealand in 1899. Almost 100 years later there was some concern about its spread. The first survey of hawthorn spread was conducted on farmland at the foot of Porters Pass, in central Canterbury (Fig. 5-6), in the 1980s by cutting a large random sample for ring counts, to allow estimating the ages of the trees. Originally, only a few hawthorn trees had been planted in 1930 near a roadside hut and across the river (green dot on the map in Fig. 5-6). The spread was observed around the original trees and the growth rate of new trees increased in the 1970s. The species spread and established further

locations, 1 km from the hut or beyond by 1970, and scattered on the hillsides, marked as "Hill" in Fig. 5-6, and along linear features, scarps and gullies, by 1985, marked as "Terrace" in Fig. 5-6. The original trees had given rise to a hawthorn population that covered about 4 km square by 2006. Two-hundred random points within 4 km square were investigated as the survey locations using a GPS. All hawthorn plants within 10 m × 10 m plots were recorded, as well as the nearest fruiting tree within 100 m in each quarter. Furthermore, an aerial photograph was used to map the landforms to cover some areas.

In this study, a two dimensional array is constructed by fixing the horizontal index x of the given array as the easting co-ordinate and the vertical index y as the northing co-ordinate of the spatial distribution map of the hawthorn distribution, then dividing it at every 100 m. The x-axis and y-axis coordinate labels run from E1 to E24 and N30 to N1 respectively, shown in Fig. 5-6, equivalent to easting from 2408800 to 2411100 and northing from 5764200 to 5767100 (details of the conversion of each coordinate is shown in Appendix 5-2). The origin of the hawthorn was found at 2410124 (E) and 5765543 (N), between E14-E15 and N16-N17 in Fig. 5-6.

Four different time scales, 1966, 1976, 1986 and 2006 (in years) of hawthorn growth patterns were investigated. All inputs were generated from populations of current hawthorn age in 2007, and the total hawthorn population at each time was estimated from the ages of trees in 2007. Thus, the total sample size (n), shown in Table 5-4, was determined for all trees that supposedly existed in 1966, 1976, 1986 and 2006, although note that the data was collected based on only mature trees, mostly at 10 years old (Kean, personal communication, 4 August, 2008).

5.8.2. Analysis of spatial weed distribution using the *K*-MSA

Details of the *K*-MSA were explained in a previous chapter (Chapter 4). This section mainly discusses the application method of the *K*-MSA and how the weight parameter setting (w) of the *K*-MSA helped the spatial analysis. The *K*-number of maximum subarrays are detected from the two-dimensional array with x-y coordinates representing easting and northing, and values (z) representing the population of hawthorn in the single 100 m × 100 m cell. The first subarray (K=1) detects the largest sum of the population of hawthorn (S-value) above the w-value (set by the user), and will be described in detail in the next section. Each subarray was ranked by the sum of hawthorn population (S), which quantifies how many hawthorn trees were detected inside the subarray. Further, the S-value is converted into the S-population value (Sp) to indicate the relative hawthorn population within a single cell, to compare the density of hawthorn in the region. Four periods of data were investigated by the

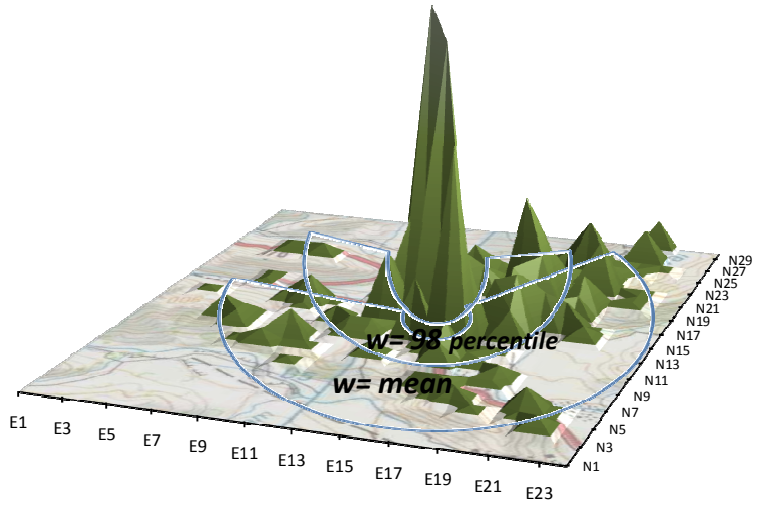


Fig. 5-7 Demonstration of different *w*-values for the *K*-MSA application.

The hawthorn population is taken from 2006 (w = 0.73 for mean, w =7 for 95 percentile). The height of each peak indicates the population of hawthorn.

K-MSA and results are reported for up to a maximum of K=6 subarrays; later subarrays are generally below the *w*-value. Each maximum subarray is detected uniquely and does not overlap with any other subarray; as the so-called *disjoint K*-MSA was applied.

5.8.3. Weight parameter (*w*) setting

The weight (w) parameter setting was developed to capture specific and more detailed aspects of maximum subarray results from flatter data sets, i.e., with no specific obvious structures in the data, to identify different subarrays over the two-dimensional array, by setting the threshold to change the detection sensitivity (Fukuda and Takaoka 2007). In this study, the hawthorn data contained many zero values (no hawthorn) over the space. The *w*-value is a threshold value, taken from measurements of z-values (the population of hawthorn in each cell), which highlights significantly high populations above the *w*-value by subtracting the *w*-value from the array (see details in Chapter 4).

A primary experiment was carried out to test several different *w*-values, e.g., mean, 50, 60, 65, 75, …, 95 and 98 percentile. The mean and 98 percentile values detected general trends in maximum hawthorn distribution and gave the most feasible and the sensitive results, capturing the extreme aggregation centres of hawthorn populated regions or positions, respectively.

Fig. 5-7 demonstrates the idea of *w*-values over the array (space). The position of 98 percentile (w = 7) cuts through the extreme top hawthorn populated regions (top 2 percentile), whereas the mean value covers most of the peaks by detecting broader and general aggregation patterns over the space. All investigations were carried out with two weight

values, mean and 98 percentile, and results were compared over different periods to help understanding and quantifying hawthorn aggregation patterns.

5.8.4. Randomization simulation tests

A simulation study is separately carried out to calculate the statistical significance of the observed maximum subarray regions (in their sizes), testing if they are random events or not. In other words, this assesses the chance observed aggregations of hawthorn are a result of hawthorn having a random spatial pattern. The randomization test used changed the x and y coordinates randomly 10,000 times to repeatedly generate random distributions of hawthorn. These random distributions of hawthorn will have the same cell-values, but the positioning of the cells will be different from the observed pattern. Each simulated set was then re-run with the K-MSA. The size of each subarray (A) was examined to locate the six maximum subarrays (K=6). Then, the cumulative sum of the areas from the first and second largest subarrays (at K=1 and 2), up to the last subarray at K=6, was separately calculated for both observed and simulated sets. The mean and standard deviation values of the simulated subarray sizes (areas, A) were calculated. An index of weed patchiness (I) was calculated from the observed maximum subarray of area (A) divided by the mean of the simulation result. A large index value indicates that the observed aggregation is larger than aggregations from the simulated random pattern. The aggregated clustered regions' significance was measured by the percentile position of the observation value in all simulated values. Very small or large percentile positions for the observed result suggest that the observed aggregation pattern, did not occur randomly. Additionally, the mean value of the simulation was compared with the observed value using a one sample mean z-test.

5.8.5. An exploratory comparison to the clustering method

The performance of the K-MSA was compared to an existing clustering method designed for ecological spatial data, Spatial Analysis by Distance IndicEs (SADIE), developed by Perry (1995) and Perry et al. (1999). SADIE uses a similar method to test for departures from randomness, based on Voronoi tessellations, but incorporating a biological model. SADIE detects clusters in the form of patches, comprising several nearby large counts, and gaps, comprising several nearby small counts, to quantify the spatial pattern in two-dimensional mapped data. Clustering regions are measured by the degree to which the unit contributes to clustering (quantified by the flows of individual points, index v_i, from *donor* sample units (greater than average abundance) and receiver units (less than average abundance). Generally, the larger index v_i value indicates patchiness, e.g., $v_i \cong 1.5$, a larger negative v_i

value indicates a gap, and v_i close to unity indicates a random placement of that unit in relation to others nearby. Additionally, SADIE provides an index (I_a) to indicate the aggregation pattern by permuting the observed set of counts amongst the sample units. The I_a value is calculated by the observed value divided by the mean value from several hundred such randomisations; when the I_a value is 1, it indicates randomly arranged counts, whereas when larger than 1, it indicates aggregation of observed counts into clusters. Further description is provided in Perry (1995) and Perry et al. (1999).

5.9. Results and discussion

Firstly, results of the *K*-MSA using the mean and 98 percentile as weight values for four periods, 1966, 1976, 1986 and 2006, will be discussed to show how the *K*-MSA identifies the generalized and the extreme centres of maximum aggregated hawthorn regions. Note that weight parameter values of the mean and 98 percentile were shown earlier in Table 5-4. Secondly, results of the randomized test (on maximum aggregated regions using w = 98 percentile) will be discussed to show how the detected aggregation areas are occurred as random events or not. Lastly, the comparison with SADIE results will be discussed.

5.9.1. Maximum aggregation of hawthorn distribution above average spread

The position of each subarray detected by the *K*-MSA with the mean weight value is drawn on the studied site geographical map for 1976, 1986 and 2006, shown in Fig. 5-8. This investigation helps identifying how the aggregated hawthorn regions and their sizes changed over four different studied periods. The number inside the subarray box in Fig. 5-8 indicates the *S*-value, which is a sum of hawthorn population detected above the mean of the total array. The first maximum subarray at *K*=1 has the largest *S*-value, *S*1, and so on. Detailed outputs of observed *S*-values, the total of the maximum subarray cell or area (*A*), their direct output coordinates (E_1, N_1; E_2, N_2) are shown in Appendix 5-3. However, all results were reported at (E_1, N_1), (E_2+1, N_2+1) to be convenient. For example, when the *S*-value is 5 and the direct output coordinate is expressed as (E_1, N_1; E_2, N_2) over one cell (100 m × 100 m), but it is represented as (E_1, N_1; E_2+1, N_2+1), this indicates that 5 hawthorn trees were observed within one cell at $E_1 \leq E < E_2+1$, $N_1 \leq N < N_2+1$.

Generally, the mean weight parameter results detect the overall coverage of maximum aggregation patterns (in Fig. 5-8). All first maximum subarrays (indicated as *S*1 in Fig. 5-8) are detected mostly within the same region (E11, N11; E21, N25) and covered the origin (E14, N17; E15, N18), marked with a star in Fig. 5-8. This may suggest that generally highly aggregated hawthorn regions or their positions were not dramatically changing over time, but

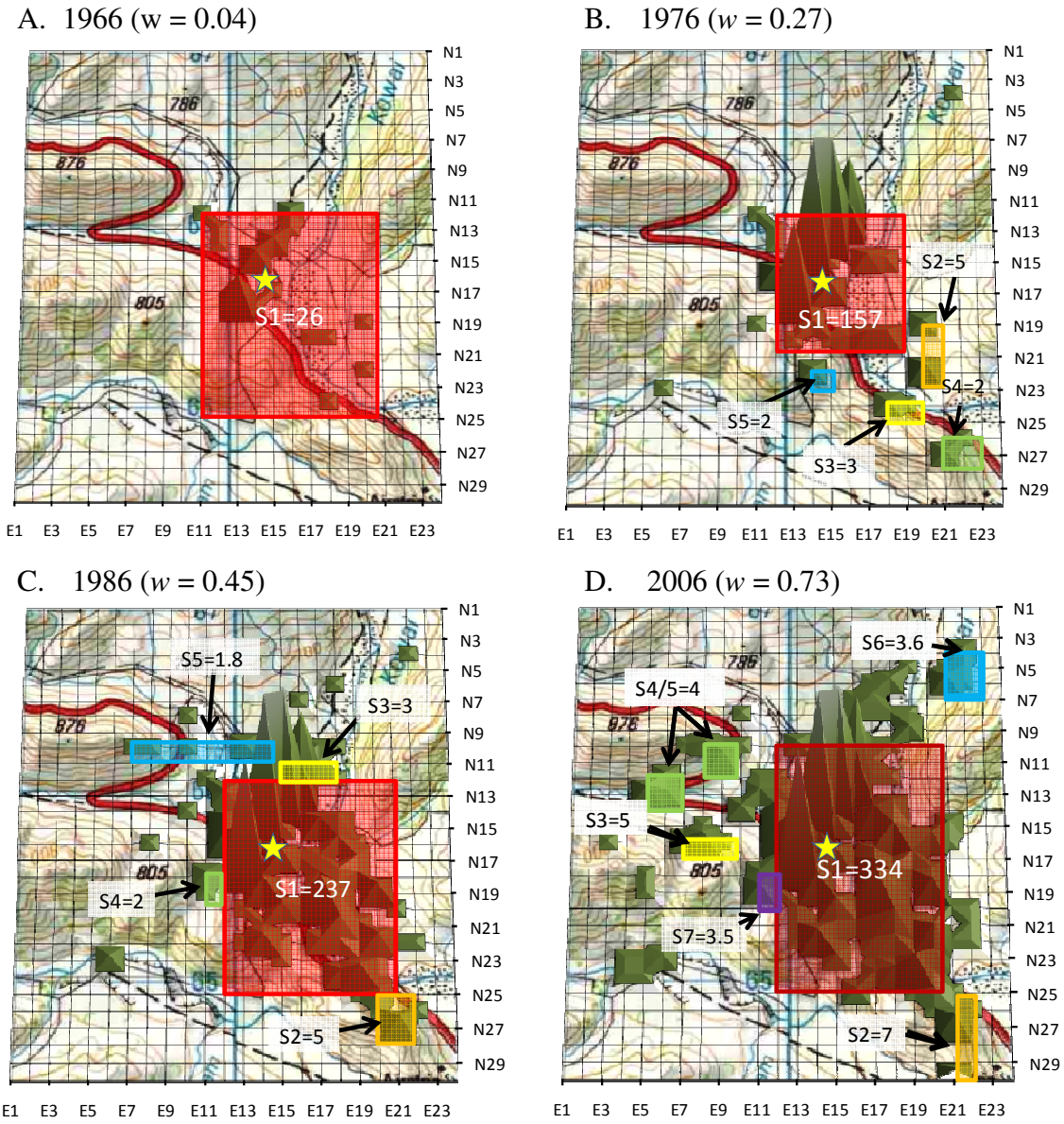


Fig. 5-8 Maximum aggregation of hawthorn populations above (w = mean) detected by the *K*-MSA.

Star marks indicate the original hawthorn location (E14, N16; E15, N17). Note that a value inside the bracket indicates each *w*-value.

the population within this region was dramatically increasing; *S*1=26 for 1966, *S*1=157 for 1976, *S*1=237 for 1986 and *S*1=334 for 2006.

The distribution of hawthorn was known to be uneven, and growth rate and suitability of habitats change over time (Williams, personal communication, 4 Aug, 2008), although the hawthorn growth also favours certain unchangeable factors such as the landscape of this region (discussed in a later section). It could be possible that the average trends of most favoured aggregated regions may not alter the fundamental aggregation patterns and positions dramatically through time. If the detected *K*=1 position can suggest the edge of the most

favoured regions for hawthorn growth, then focusing control within this region may have helped stop weed spread.

The *K*-MSA provides some clues to how the weed spread over time. After 1966, the next largest subarrays (*K*=2, …, 6) detected numbers of new but highly aggregated hawthorn regions outside of this *K*=1 region. Following the order of larger to smaller *S*-value suggests generally maximum aggregated regions favoured firstly the south-east direction towards the lower Kowai river in 1976 (*K*=2, …, 5), then in 1986 (*K*=2), the north-east direction towards the upper Kowai river in 1986 (*K*=3) and in 2006 (*K*=6), and eventually the spread reached the north-west direction in 1986 (*K*=5) and in 2006 (*K*=3, 4, 5, 7).

The spatial spread is dictated by the availability of roost sites for the deposition of hawthorn seeds by blackbirds and for the germination, growth and survival of the resulting bushes. The underlying pattern causing this roost site availability is dictated by topography, soil conditions, and land use changes through time (Williams, personal communication, 5 Aug, 2008). The establishment of hawthorn was closely related with the presence of the indigenous spiny shrub, matagouri (*Discaria toumatou*). Early in the invasion of hawthorn, the matagouri was the only place for blackbirds to nest. Matagouri facilitates hawthorn by acting as roost sites and blackbirds to perch and to defecate. Their dense prickles exclude grazing which allows the resulting seedlings to grow to hawthorn (Williams, personal communication, 4 Aug, 2008). Generalized aggregation spatial patterns found were that hawthorn firstly aggregated near the origin. Following the aggregates spread along the Kowai river (towards both the upper and lower Kowai river), and north-west (far and independent regions from the origin). These results may possibly help identifying patterns of roost site locations for the future weed management and control process.

The first maximum region in 1966 (Fig. 5-8, A) covered most of the hawthorn populated regions; this was due to generally low hawthorn populations in 1966 that were not significantly high enough (above the mean) to separate into individual aggregated regions. However, these low populated regions at the bottom of the Kowai river in 1966 became highly populated in 1976. Hence, the previous first maximum subarray region in 1966 separated to form smaller *K*=1 areas in 1976 (Fig. 5-8, B). In 1986 (Fig. 5-8, C), these several small aggregated regions (K=2 to 5) at the bottom of the Kowai river further increased in population since 1976 to fill up the gap between the *K*=1 region. This produced a very large aggregated region in 1986 (*S*1=237). Further, it continued to increase density within the same region, by 2006 (*S*=334), in Fig. 5-8, D.

The *K*-MSA results using the mean weight value provide information on how the maximum aggregated regions changed to form or separate its aggregation over time. However, outputs using the mean weight value only identify the general trends, and are not sensitive enough to detect how the maximum aggregated regions inside the clustered regions were changing over time. Hence, the use of the 98 percentile weight value detects more specific position of the centre of maximum aggregated regions.

5.9.2. Maximum aggregated hawthorn distribution pattern above 98 percentile spread

Investigation of the maximum subarray analysis using a 98 percentile weight value provides information on the centre position of the maximum aggregated regions. Similar to

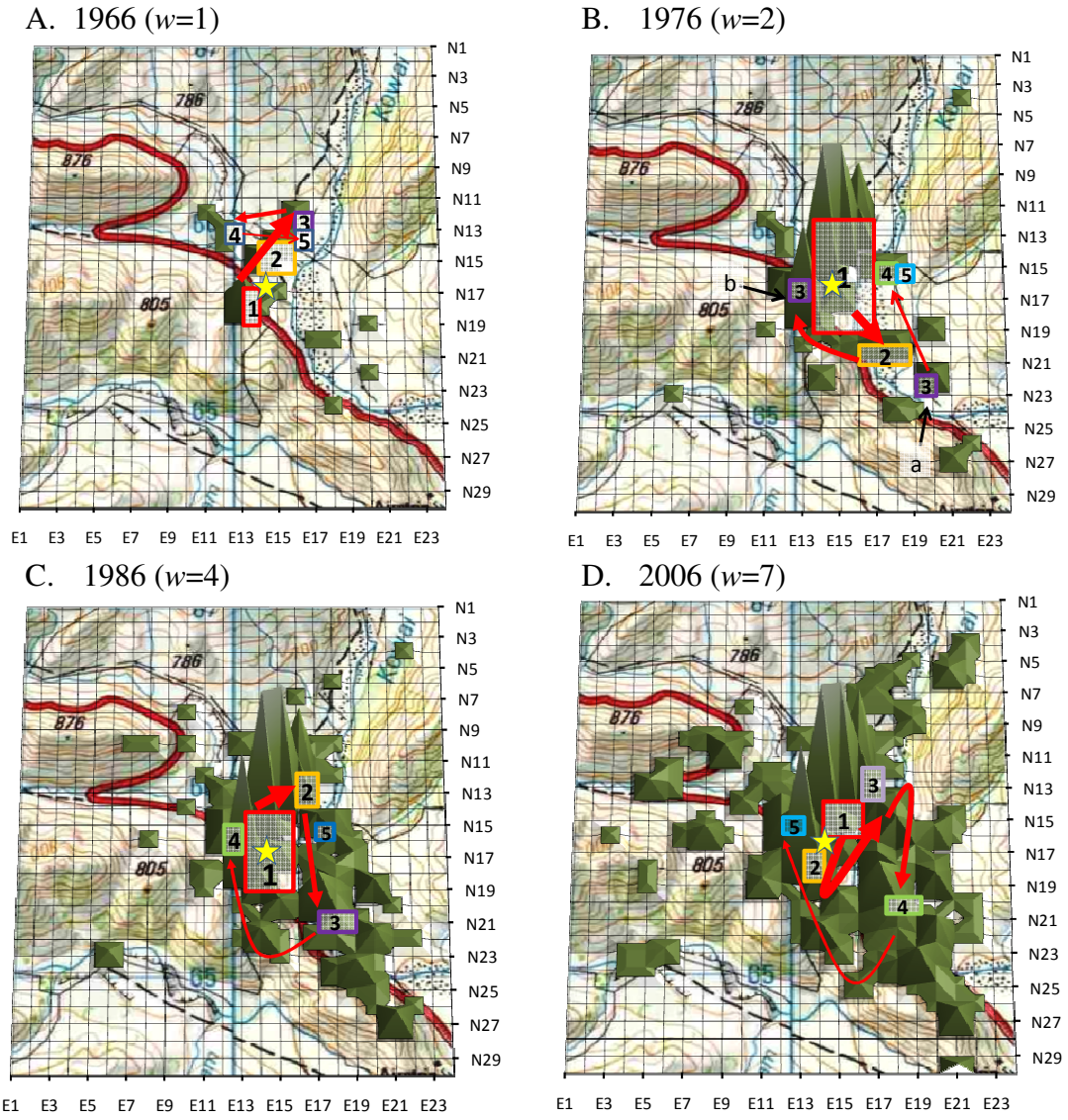


Fig. 5-9 Maximum aggregation of hawthorn populations above (*w*= 98 percentile) detected by the *K*-MSA.

Original hawthorn location marked with a star.Note that each value inside the bracket indicates *w*-value.

the previous analysis, the detected maximum subarray positions are shown on the map in Fig. 5-9. The number in the box indicates *K*-value (not *S*-value) and the arrow indicates the direction of the shifts in maximum subarray from large to small *K*-value (equivalent to the order of *S*-value). Note that detailed outputs are shown in Appendix 5-4.

Firstly, the centre of maximum aggregated regions was investigated by observing visible changes of the maximum subarray regions on the map in Fig. 5-9. Secondly, detailed observations were carried out to examine the relative density of hawthorn proportion (above top 2 percentile) over a single cell (100 m × 100 m) to compare the density among different maximum aggregated regions over time. Thirdly, the detailed coordination or position of maximum aggregated regions was examined to show how the centre of maximum hawthorn aggregation regions changed or shifted their position over time.

5.9.2.1. Spatial distribution pattern of maximum aggregated regions (*w* = 98 percentile).

Generally, all maximum subarrays identified using a 98 percentile weight value were smaller than these with the mean weight value because the algorithm identifies the top 2 percentile of the total hawthorn population of the array. In 1966 (Fig. 5-9, A), the highest maximum aggregated region was very small and its shape was elongated from north to south along the road (red line on the map) in the hill site (*S*1=10, *A*=2 at E13, N17; E14, N19) adjacent to the original hawthorn location. This shape is similar to the shape of the mean weight maximum subarray. Note that *A*=2 includes two cells, 100 m × 200 m. The second maximum subarray was found from the terrace site just above the original hawthorn but its aggregated region size is larger (*S*2=3, *A*=4, E14, N14; E16, N16) than the first maximum subarray. The third maximum subarray was found over a very small region (*S*3=1, *A*=1, E16, N12; E17, N13) further north-east by the upper Kowai river.

In 1976 (Fig. 5-9, B), all small aggregated regions in 1966 became highly populated enough to be detected with a 98 percentile weight value. Previously scattered aggregated independent regions in 1966 filled the gaps between them to form the largest aggregated region with much higher hawthorn population over a larger area (*S*=93, *A*=28, E13, N12; E17, N19) at the terrace site. Two new medium aggregated regions were detected (*S*2=4, *A*=4, E16, N20; E19, N21 at *K*=2, and *S*3=2 for *A*=2, E12, N16; E13, N17) towards the south-east, lower Kowai river. Additionally, further smaller aggregated regions (K=3b, 4) were found just next to the largest aggregated region (*K*=1). Generally, positions of maximum regions were similar using the mean weight. This suggests that the observed distribution

pattern in 1976 was possibly simpler (as either examining the general or the centre of the maximum aggregation pattern gives similar results) rather than complex.

In 1986 (Fig. 5-9, C), the maximum aggregation positions and their directions were generally similar to 1976, but with increased hawthorn population (*S*-value). However, the highest maximum aggregated region at *K*=1 became smaller (*S*1=80, *A*=15, E13, N14; E16, N19) over the origin, indicating a denser spread. The second maximum aggregated region (*S*2=25, *A*=2, E16, N12; E17, N14) was detected from the north-east side of the original hawthorn region towards the upper Kowai river. The third maximum aggregated region (*S*3=9, *A*=2, E17, N20; E19, N21) was detected towards the south-east lower Kowai river, the previous location of the second maximum aggregated region in 1976.

In 2006 (Fig. 5-9, D), the hawthorn population had become significant enough to allow the *K*-MSA to detect many more aggregated centre points (*K*-value up to 5 is shown in Fig. 5-9, but at least 8 clustered points were detected), suggesting that aggregation patterns became more heterogeneous and complex. Note that this aggregate pattern (the large *K*=1 region) was not observed when using the mean as the weight value. Interestingly, the highest maximum aggregated region (*K*=1) was not detected over the hawthorn origin. The location of the *K*=1 region was now shifted to the north-east of the original hawthorn region, towards the terrace site (*S*1=56, *A*=4, E14, N14; E16, N16). The origin of hawthorn was now detected between the *K*=1 and *K*=2 subarrays. The second maximum aggregated region was detected just above the origin, in the hill site along the road, covering a small area (*S*2=43, *A*=2, E13, N17; E14, N19). No specific subarray regions were detected to include the origin in 2006. The third maximum aggregated area (*S*3=27, *A*=2, E16, N12; E17, N14) was detected towards the upper Kowai river where the second maximum aggregated region was previously. Then, the fourth region (*S*4=8, *A*=2, E17, N20; E19, N21) was detected at the previous location of the third maximum subarray region and so on.

Generally, positions of maximum aggregated regions did not alter over time between 1966 and 2006, except that the hawthorn origin was no longer a part of the highest maximum aggregated region in 2006. One explanation is that the previously aggregated regions have reached full capacity, and were no longer available for its further growth.

5.9.2.2. Evolution of the maximum aggregated hawthorn positions.

To summarize the above findings, each coordinate (E_1, N_1; E_2, N_2) of the maximum subarray position (*K* up to 5), *S*-population value (above 98 percentile, *Sp*), and areas of the maximum subarray regions (*A*), are shown in Fig. 5-10 for detailed investigation. Each

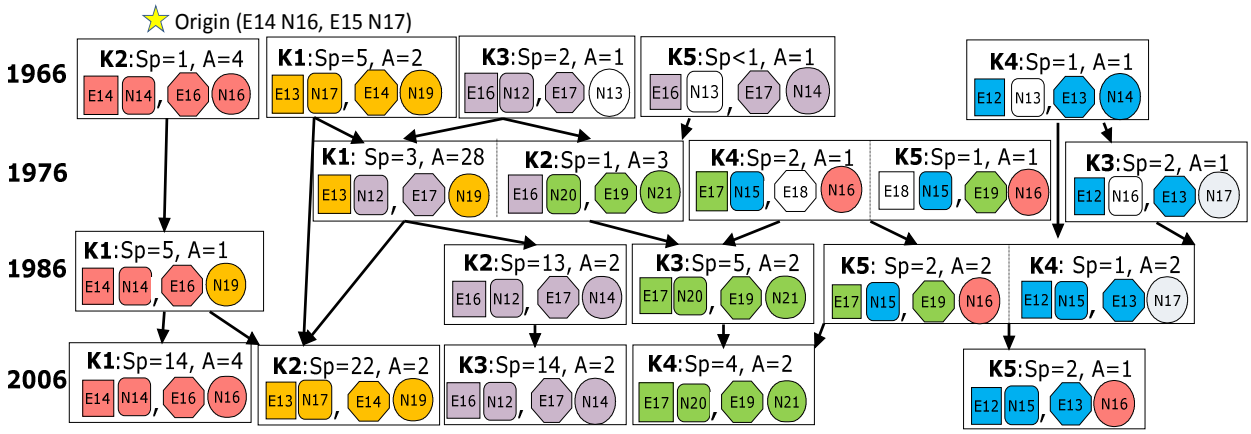


Fig. 5-10 Summary of the maximum aggregated hawthorn regions (w = 98 percentile).
The same coloured cells indicate the exact same coordinate of maximum subarray region.

identical colour in Fig. 5-10 indicates the same coordinate respectively for each position of E_1, N_1; E_2, N_2.

This display assists in determining how the position of each aggregated hawthorn regions evolved to shift or grow its population. From examining the exact position of the maximum aggregated hawthorn region centres in Fig. 5-10, the original hawthorn aggregated regions in 1966 were not significantly altered over time. Most of the coordinates that were selected in 1966 (red, orange, purple and blue cells in Fig. 5-10) were also selected to form later aggregated regions, but with increasing density of hawthorn populations over time, i.e., the *Sp* values were increasing each period. Prior to 2006, a few non-overlapped coordinates were observed, shown as white cells in Fig. 5-10. However, all coordinates were eventually filled up with previously selected regions in 1966 and 1976 by 2006, i.e., no white cells are left in 2006. As previously discussed, the second maximum subarray (K=2), hill site, in 1996 had developed into the largest aggregated region (K=1) since 1986, in Fig. 5-10.

Interestingly, the third maximum subarray position (K=3, north-east towards the Kowai river) in 1966 was developed into two future aggregated regions from the one point of its coordinate; E16, N12 (purple cells in Fig. 5-10); to the most aggregated region in 1976 (K=1, largest aggregated region covers the origin and hill site) and to the second maximum aggregated hawthorn region in 1976 (K=2, south-east). Further, this coordinate was continuously selected to form the second maximum aggregated hawthorn region in 1986 (K=2, north-east, near the upper Kowai river) and the third maximum aggregated region in 2006 (K=3, north-east, near the upper Kowai river). Another interesting observation found from Fig. 5-10 is that the new coordinates, E17, N20, E19, N21 (in green cells) appeared in 1976 and these coordinates were detected again as K=3 in 1986 and K=4 in 2006.

5.9.2.3. Density comparison of the maximum aggregated regions

The density comparison of the detected top two percentile of the total array for the four periods is summarized in Fig. 5-11. Density of hawthorn proportion (top two percentile) over one cell (100 m × 100 m) is defined by the *S*-proportion (*Sp*), shown as each value in Fig. 5-11. The *Sp*-value is calculated by dividing the *S*-value by the area (*A*) to allow comparisons of the hawthorn density over different *K*-regions among four periods.

The later periods generally have higher *Sp* values across different *K*-regions over time (Fig. 5-11). In particular, the *Sp*-value at *K*=1 in 2006 was approximately tripled (*S*=14) compared with 1966 (*Sp*=5), 1976 (*Sp*=3) and 1986 (*Sp*=5). However, the highest *Sp*-value (*S*=22) was detected from the second maximum subarray (*K*=2) in 2006. By combining results from the previous section and this analysis, the different aggregation pattern between the hill site along the road (across the road of the origin) and the north-east terrace site (above the origin towards the Kowai river) is highlighted. The position of the highest aggregated region (*K*=1) has shifted from covering the origin region to the terrace site in 2006, since the highest top two percentile population was observed (*S*1=56). However, the terrace site shows the lower density (*Sp*=14, *K*=1 in Fig. 5-11), whereas the hill site shows lower top 2 percentile population (*S*2=43) than the terrace site, but has higher density (*Sp*=22 for *K*=2 in Fig. 5-11). This pattern is similarly shown from 1986, where the *Sp* value is higher at *K*=2 (located in the north-east towards the upper Kowai river) than at *K*=1 (which covered a large area over the origin), suggesting that the denser growth pattern was found from the *K*=2 region compared with the *K*=1 region, but a higher population was detected in the *K*=1 region than the *K*=2 region.

To summarise, the terrace site and hill site (in particular along the road) generally were

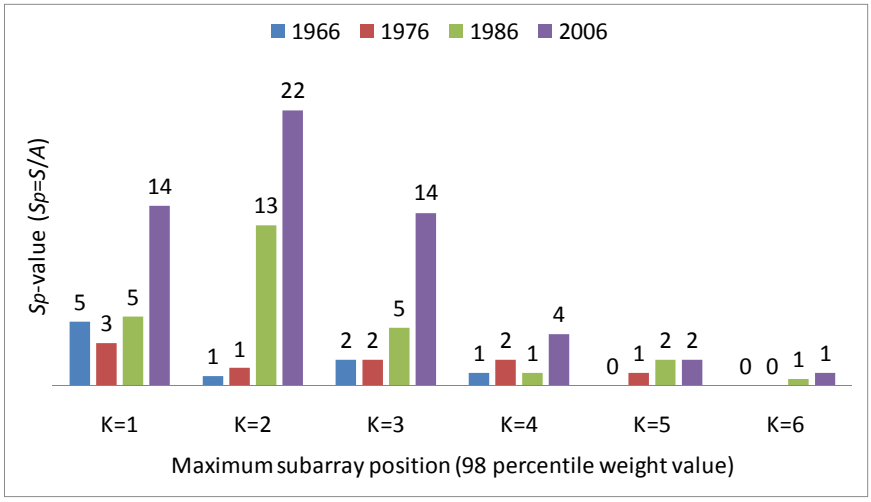


Fig. 5-11 Density comparison by *Sp*-value (*w*=98 percentile) across various maximum aggregated regions.

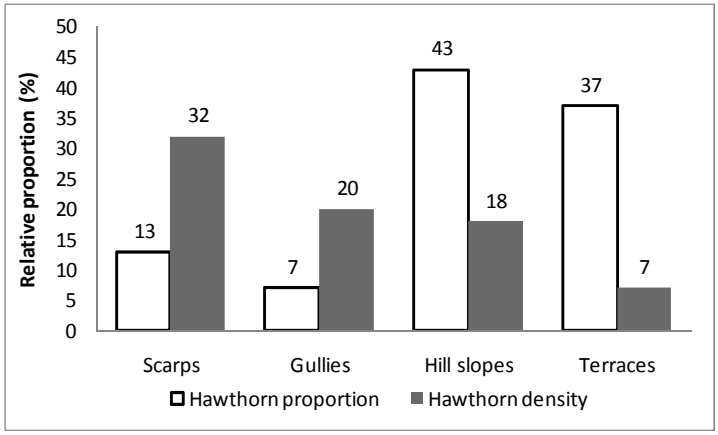


Fig. 5-12 Distribution of hawthorn population and density among different landscape.
Information was provided by Williams (personal communication, 6 Aug, 2008).

sites of high hawthorn aggregations over the 40 years. Although new spreads were observed and each detected region increased in population over time, the central maximum aggregated positions since the early age, especially the terrace and hill site, did not dramatically change over time. However, the hawthorn aggregation patterns were slightly different between the terrace and hill sites, with a more thinly spread distribution pattern for the terrace site, and an intensively grown dense pattern for the hill site, but only along the road adjacent to the origin.

From previous investigation from Williams and Buxton (1986) and expert knowledge (Williams, personal communication, 6 Aug, 2008), the road margins provided a roost site for blackbirds until the 1980s, before land was retired from grazing. Besides, the fences along the road may accumulate more seeds through droppings from birds as they perch on the fences, and the wire prevents sheep and rabbits from eating the young bush that had now become trees. The expert opinion is that the overall relative density of young and old hawthorn is similar over time, but the different landscape, scarps, gullies, hill slopes and terraces, are characterized by different degrees of hawthorn density. In spite of higher distribution of hawthorn detected from hill slopes (43%) and terraces (37%), the higher density was observed from scarps (32%), gullies (20%) and hill slopes (18%), shown in Fig. 5-12, where the road is located very close to a scarp above the river (Williams, personal communication, 6 Aug, 2008).

Here, the maximum subarray region detected has an elongated shape over the small area along the road rather than the large square region that covered the large area over the hill slopes, e.g., the K=2 and K=5 region in 2006 (Fig. 5-9, D). In contrast to this, the K=1 region above the origin in the terrace site generally shows the wider square coverage, e.g., the K=1 region in 2006 (Fig. 5-9, D). Additionally, the K=4 region in 2006, south-east towards the lower Kowai river, which was generally detected at the K=3 region in 1986 (Fig. 5-9, C) and

the *K*=2 region in 1976 (Fig. 5-9, B), was detected as an elongated maximum subarray shape from the west in the lower position of the terrace site to the east, towards to the road. Note that this region was constantly detected above 98 percentile weight values since 1976, but generally detected at lower *K*-value in later years. This indicates the population in this region is high, but not higher than other regions.

5.9.2.4. Randomization tests for observed and simulated maximum aggregated regions

All detailed outputs of the simulation test are shown in Appendix 5-4. Table 5-5 shows the percentile position of observed maximum subarray areas (cell sizes) among the simulated random distributions' maximum subarray results (*n*=10,000). Generally, all values in Table 5-5 indicated that the detected maximum subarray sizes did not occur as random events, since they had either very high or very low percentile position values (> 0.0001). The overall percentile value of the observed data ranged from 0.862 (*K*=2 in 1976) to 0.998 (K=1 in 1976) or > 0.0001, e.g., *K*=4 in 1966. The large percentile indicates that the observed

Table 5-5 Percentile position of observed maximum subarray over the simulation test*.

K- value	*K*=1	*K*=2	*K*=3	*K*=4	*K*=5	*K*=6
1966	0.917	0.995	0.000	0.000	0.000	0.000
1976	0.998	0.862	0.000	0.000	0.000	0.000
1986	0.997	0.871	0.939	0.894	0.000	0.961
2006	0.876	0.935	0.947	0.931	0.000	0.000

* p-value is 0.000 for all tests.

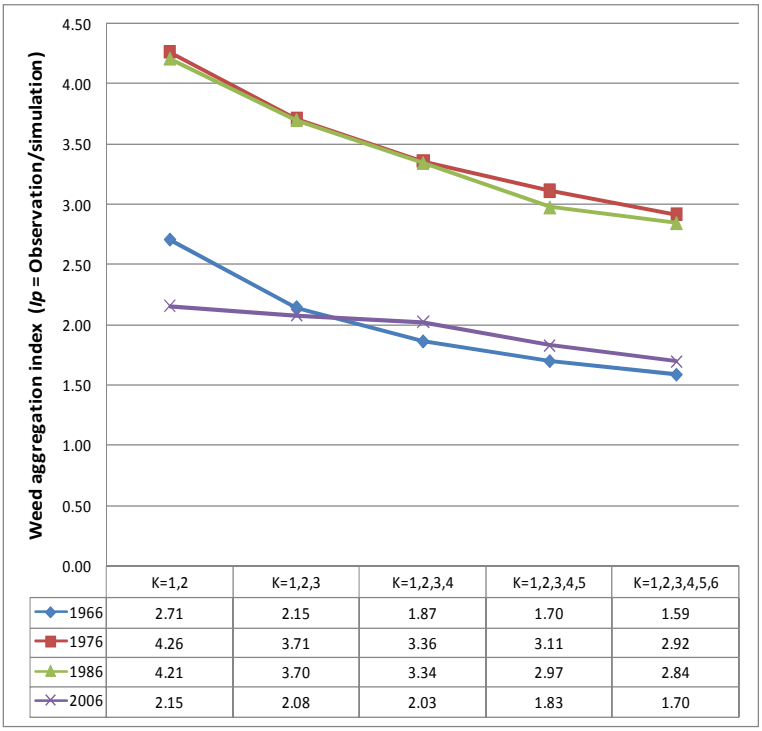


Fig. 5-13 Weed patchiness index for observed and simulated maximum subarray results.

maximum subarray areas were larger aggregates than the subarrays from the simulation and the small percentile indicates the opposite. For example, only 0.2% of the highest maximum aggregation region from a random distribution was larger than the observed array in 1976 (K=1 in Table 5-5). There were some variations in these percentile values, e.g., relatively lower percentile values are observed from K=2 in 1976 (0.862) and K=1 in 2006 (0.867).

Fig. 5-13 shows summary results of an index of weed patchiness (I), calculated from the observed areas (total cell size) divided by the mean value of the simulated maximum subarray areas using a 98 percentile weight value (note that the number of simulations is 10,000). The cumulative sum of the subarray area was calculated (up to K=6) separately for the four periods. This test investigates how large or small patchiness is observed compared with the results for the simulated (random) results. Larger aggregation index, where I is more than 1 indicates that the observed maximum aggregated area is larger than the simulated results, and I less than 1 indicates the opposite.

From Fig. 5-13, all observed results are more than 1, indicating that the detected observed aggregation was larger than the results for the simulated patterns. The highest I values were found from the combination of K=1 and K=2 regions. The I values for the later cumulative sum results (sum of up to K=6) were progressively smaller, indicating that the highly aggregated detected regions were much larger than simulated results. Adding smaller subarrays to the sum eventually formed a sum closer to the sum from the simulated populations.

Another observation from Fig. 5-13 is that similar weed aggregation index trends were found between 1966 (blue) and 2006 (purple), with lower index values, and between 1976 (red) and 1986 (green), with higher index values. This suggests that the mid periods (1976 and 1986) have larger aggregated areas than the simulation and results from the early and late periods. From the spatial distribution on the map in Fig. 5-9, the common features between 1976 and 1986 and between 1966 and 2006, respectively, are that the detected maximum subarray areas of K=1 (the origin area) in particular were larger for 1976 and 1986, but smaller for 1966 and 2006. Another explanation for the similarity between 1966 and 2006, and between 1976 and 2086 follows.

The K-MSA detected small aggregated regions with high density hawthorn populations, for both 1966 and 2006, but with different background of hawthorn populations. The spatial distribution of 1966 was dominated by empty space, thus, the maximum aggregated regions were detected where hawthorn existed (above 98 percentile) with a lower population (S-value). On the other hand, the spatial distribution of 2006 was highly occupied by hawthorn,

thus the top 2 percentile of hawthorn populations were detected for the centre of the maximum aggregated regions over highly populated hawthorn populations, giving a higher population (*S*-value). One obvious similarity between 1976 and 1986, was that observing the largest aggregated region (*K*=1) suggests large spread (Fig. 5-9) but more specific aggregated patterns (since the high weed aggregation index values were detected).

Detecting similar aggregation index values among two sets of periods may suggest: 1) populations were recognised as densely aggregated over a small region for the earlier and later periods, but they have a different background condition that is dense growth over the empty or highly populated space, and 2) the mid periods are perhaps spread over larger areas, which could be due to filling up the space between different aggregations.

5.9.2.5. Application of the clustering method, SADIE

An exploratory test of SADIE was performed to show how this clustering method, developed specifically for ecological studies provides additional information on the hawthorn distribution. Results of the aggregation index (I_a) and its significance (p-value) are summarised in Table 5-6, and show that aggregation patterns became least random towards the later periods (I_a-value is getting bigger and smaller p-value), indicating that hawthorn counts were getting more clustered over time, i.e., closer to I_a=1 indicates randomly arranged counts. The aggregation was not statistically different from random (I_a = 1.23) in 1966 and its result was not significant (p=0.1000), but the aggregation pattern became significant (p=0.029) in 1986, and hawthorn counts were found to be significant in 2006 (p=0.0004).

SADIE or other clustering methods detect the edge of aggregated or clustered regions, whereas the *K*-MSA detects independent subarrays (regions) that have maximum counts. Fig. 5-14 shows a contour map drawn by the SADIE cluster index. Larger cluster index values (>1.5) indicate patchiness, large negatives indicates gaps, and values close to unity indicates a random placement of that unit in relation to others nearby (Perry et al. 1999). SADIE detects each position of patchiness (Fig. 5-14), as to help identifying the location of all clustered region of the average abundance. The widths between contours were becoming narrower over time, indicating that various levels of aggregation are observed (including random ones and gaps), suggesting the spatial aggregation became more complex over time.

Table 5-6 Assessments of SADIE results.

Studied period (year)	1996	1976	1986	2006
Aggreagation Index (I_a)	1.23	1.35	1.69	2.11
p-value	0.1000	0.0447	0.0029	0.0004

As SADIE detects the average abundance, results of SADIE, indicating the patchiness (the region inside the dark yellow for the cluster value >1.5 in Fig. 5-15) and results of the highest maximum

subarray region (K=1) from the K-MSA using the mean weight value detected similar aggregated locations, in Fig. 5-15.

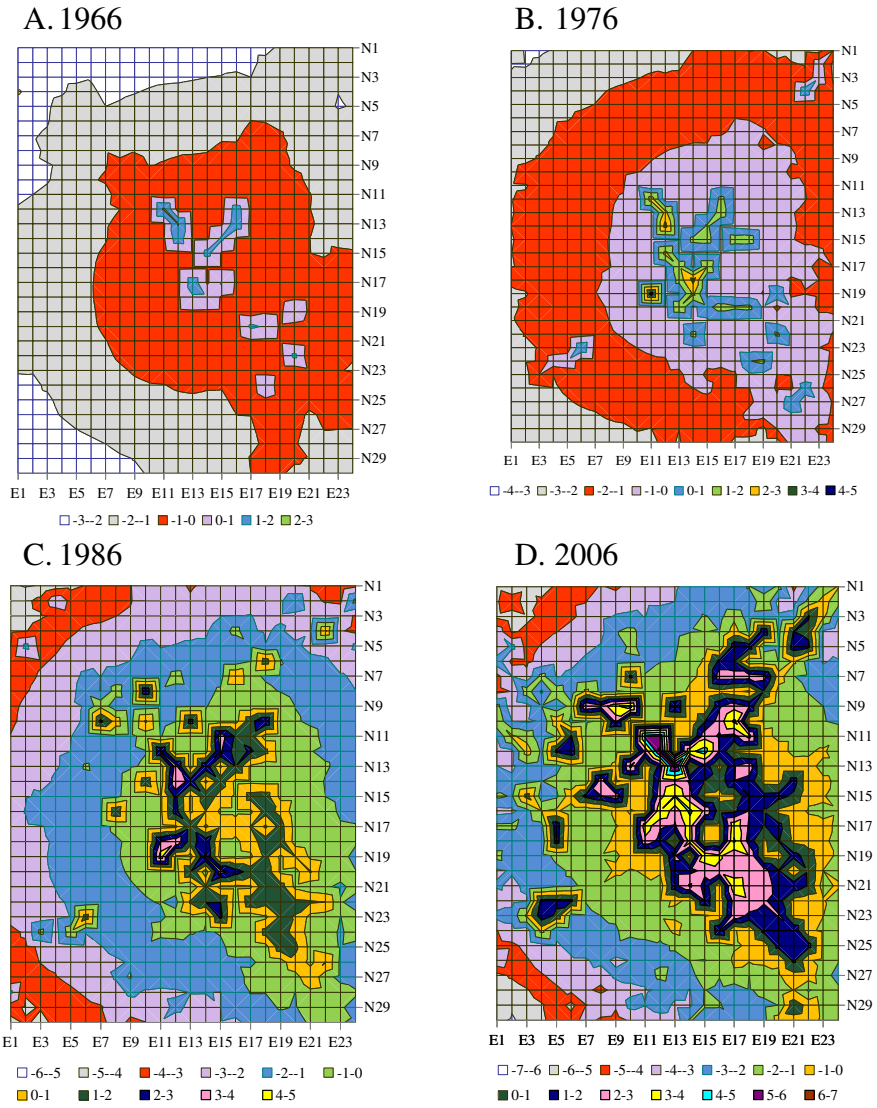


Fig. 5-14 Contour plots from SADIE clustering results.

Larger values (about 1.5) indicates patchiness, large negative values (<-1.5) indicates membership of a gap, values close to unity indicate a random placement of that unit in relation to others nearby.

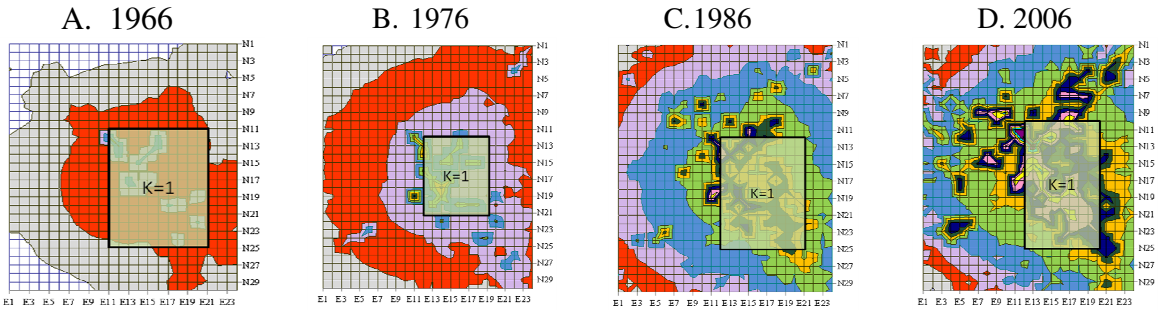


Fig. 5-15 Comparison of the largest aggregated regions detected by the K-MSA using mean and SADIE.

The patchiness is indicated inside the dark yellow line, where were overlapped with K-MSA K=1 region.

SADIE investigated the average abundance over the space. SADIE detects the outside or edge of aggregated regions and the *K*-MSA detects the general positions and centres of the maximum aggregated regions. Both methods have advantages over each other, and combining or comparing the two, the results would provide more advanced information to understand the spatial distribution of hawthorn.

5.10. Conclusions

This study introduced the use of the new *K*-Maximum Subarray Algorithm (*K*-MSA) to help understanding the spatial hawthorn distribution pattern and the change in this pattern over time. The *K*-MSA can be used to detect the generalised positions and centres of maximum aggregated hawthorn regions by changing the weight parameter to the mean and 98 percentile of the total array. Generally, several new aggregations were formed over time outside the most commonly detected aggregated regions, in which positions were not dramatically changing but the population within these maximum aggregated region increased over time. The centre of most aggregated regions was generally detected from the hill and terrace site above or including the origin site. This analysis suggests that the highest aggregated regions were similarly found from the early (1966) and later (2006) stages of hawthorn growth; the small area along the road in the hill site to indicate the dense distribution, and over the large area in the terrace site to indicate the wide spread. Note that these regions were specifically detected as the highest aggregated regions but with different background hawthorn distribution patterns; over generally empty space for the early stage and over generally high hawthorn distribution for the later stage. The hawthorn distributions during the mid periods (1976 and 1986) were detected to fill up the space between the highly aggregated regions detected in the early and later stage, indicating intermediate distribution patterns and wider spreads over a larger area than the highly aggregated spots over the small regions. Detected results were similar to expert knowledge, but this study helps quantifying its mechanism.

Furthermore, the randomisation test suggests that the detected aggregation regions (in their size) were not random events and the detected position of the generalised maximum aggregation region (the larger region over the origin) from the *K*-MSA was similarly detected from another clustering method, SADIE. A combination of the *K*-MSA and another clustering technique can provide different types of knowledge to help the weed management and control: the *K*-MSA detected the mechanism of inside the most aggregated regions, whereas the clustering method detects the outside of the aggregated region by detecting its edge.

At the moment, the algorithm is written in the C programming language and it may not be easy for non C programmers to freely try the method. Also, direct outputs were not designed to be applicable for the spatial analysis, e.g., outputs are not plotted. In the near future, the *K*-MSA will be implemented as a part of open source GIS software, such as SAMT, developed by Wieland et al. (2006), to be more user-friendly (will be introduced in Chapter 7). Also, currently the *K*-MSA only detects the edge of cells, but to be more competitive to clustering algorithms, the further development of the *K*-MSA allows detecting the most aggregated regions with softer edges.

5.11. Acknowledgement

Thanks to Dr. P Williams (Landcare Research, Nelson) and Dr. J Kean (AgResearch, Lincoln) to provide me the studied data and knowledge.

5.12. References

Bae SE, Takaoka T (2006) Improved algorithms for the *K*-Maximum Subarray problem. Comput J 49:358-374.

Bae SE, Takaoka T (2007) Algorithms for *K*-Disjoint Maximum Subarrays. Int J Found Comput Sci 18:319-339.

Dietz H, Edwards PJ (2006) Recognition that causal processes change during plant invasion helps explain conflicts in evidence. Ecol 87:1359-67.

Fukuda K, Takaoka T (2007) Investigation of the maximum association for suicide rate and social factors using Computer Algorithm, In Oxley L and Kulasiri D. (eds) MODSIM 07, 1381-1387.

Hess D, van Lieshour, M-C, Payne B, Stein A (2001) A review of spatio-temporal modeling of quadrat count data with application to striga occurrence in a pearl millet field. Int J Appl Earth Observ Geo-Inform 3:133-138.

Perry JN (1995) Spatial Analysis by Distance Indices. J Anim Ecol 64; 303-314.

Perry JN, Winder L, Holland JM, Alston RD (1999) Red-blue plots for detecting clusters in count data. Ecol Let 2 :106-113.

Swetnam RD, Mountford JO, Armstrong AC, Gowing DJG, Brown NJ, Manchester SJ, Treweek JR (1998) Spatial relationships between site hydrology and the occurrence of grassland of conservation impotence: a risk assessment with GIS. J Environ Manag 54: 189203.

Wieland R, Voss M, Holtmann X, Mirschel W, Ajibefun I (2006) Spatial analysis and modeling tool (SAMT): 1. Structure and possibilities. Ecol Inf 1:67-75.

Wang J, Christensen *S*, Hansen PK (2008) A method for building spatial model of annual weed seed dispersal from experimental data and its application to simulating Bromus sterilis population dispersal. Ecol Model 210:446-452.

Williams P, Buxton RP (1986) Hawthorn (*Crataegus monogyna*) populations in mid-Canterbury. NZ J Ecol 9:11-17.

Williams P, Buxton R, Ferris *S*, Kean J (2007) Modelling the spread of hawthorn in mountain Canterbury, NZIMA programme on Modelling Invasive Species and Weed Impact, Christchurch, 16-20 April (presentation slides are be available by contacting Dr. Williams at Landcare Research Ltd, Nelson).

5.13. Appendices

Appendix 5-2 Lists of each coordinate for the geographical maps.

Northing	N	Northing	N	Easting	E	Easting	E
5767100	N1	5765600	N16	2408800	E1	2410300	E16
5767000	N2	5765500	N17	2408900	E2	2410400	E17
5766900	N3	5765400	N18	2409000	E3	2410500	E18
5766800	N4	5765300	N19	2409100	E4	2410600	E19
5766700	N5	5765200	N20	2409200	E5	2410700	E20
5766600	N6	5765100	N21	2409300	E6	2410800	E21
5766500	N7	5765000	N22	2409400	E7	2410900	E22
5766400	N8	5764900	N23	2409500	E8	2411000	E23
5766300	N9	5764800	N24	2409600	E9	2411100	E24
5766200	N10	5764700	N25	2409700	E10		
5766100	N11	5764600	N26	2409800	E11		
5766000	N12	5764500	N27	2409900	E12		
5765900	N13	5764400	N28	2410000	E13		
5765800	N14	5764300	N29	2410100	E14		
5765700	N15	5764200	N30	2410200	E15		

Appendix 5-3 Outputs of observed and simulated maximum subarrays (w = mean).

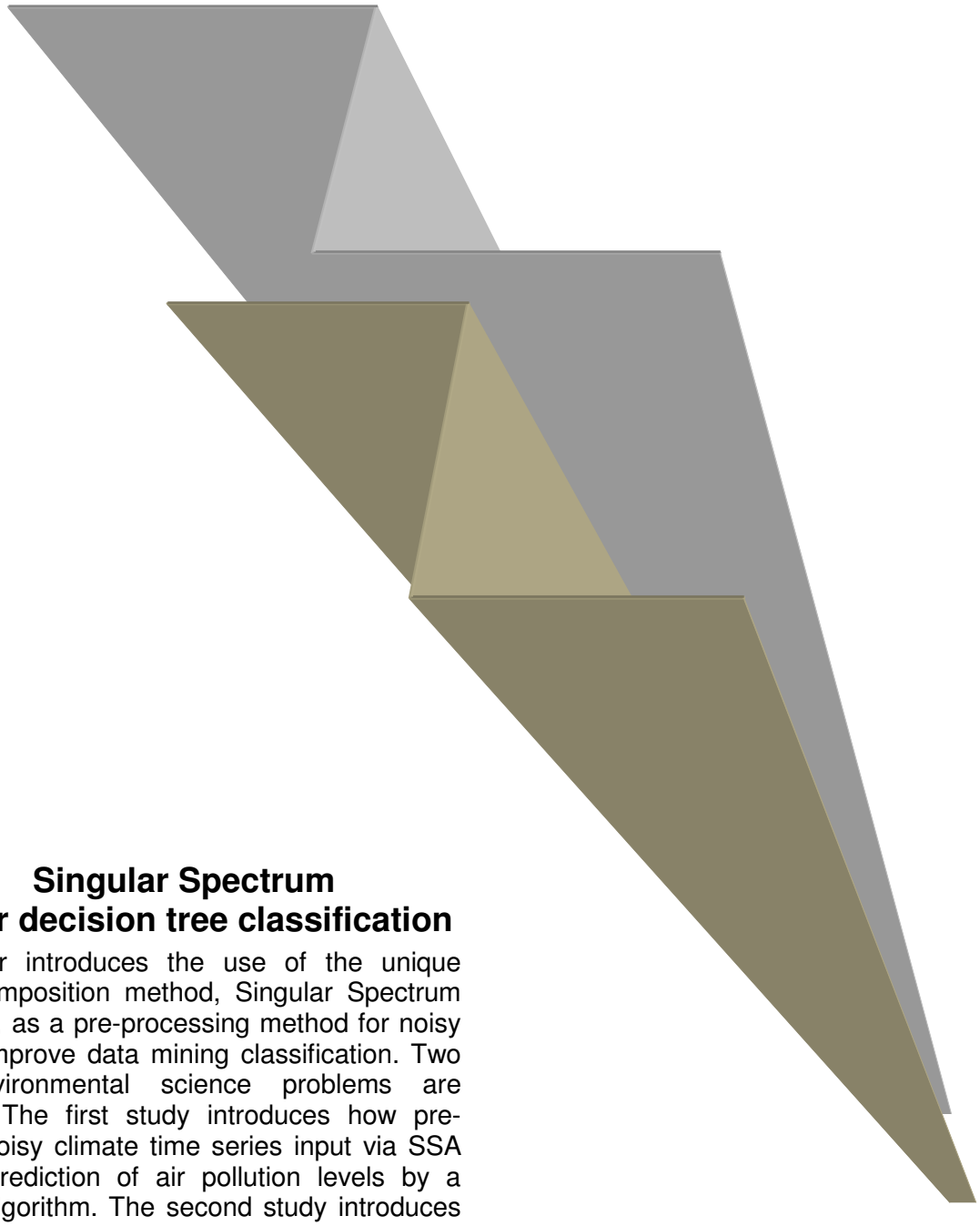
	1966						1976						1986						2006					
K=1	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	130	26.22	11	12	20	24	63	157.20	12	12	18	20	117	237.03	12	12	20	24	135	333.94	12	10	20	24
Mean	111.39	14.95	7.82	9.74	17.16	21.20	148.03	85.49	7.21	8.91	17.85	22.03	160.50	101.03	6.85	8.61	17.95	22.36	182.21	111.15	6.61	8.18	18.44	22.88
SD	69.80	2.02	5.80	7.39	5.86	7.41	84.17	15.29	5.43	6.92	5.43	6.89	85.39	18.27	5.30	6.74	5.44	6.77	92.36	21.01	5.34	6.69	5.27	6.66
Percentile*	0.648	-	0.7	0.64	0.56	0.53	0.164	-	0.78	0.69	0.39	0.31	0.354	-	0.8	0.7	0.51	0.47	0.348	-	0.81	0.64	0.47	0.44
Max	440	22	24	30	24	30	468	138	24	30	24	30	483	166	24	30	24	30	529	197	24	30	24	30
K=2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	1	-0.04	1	1	1	1	4	4.93	20	19	20	22	6	5.28	20	25	21	27	6	6.64	21	25	21	30
Mean	29.91	3.30	10.04	12.67	14.77	18.14	18.62	23.02	10.81	13.57	14.15	17.46	26.94	30.49	10.31	13.07	14.91	18.25	37.93	39.68	9.77	12.15	15.37	18.67
SD	27.28	1.69	7.22	9.06	7.25	9.13	22.49	14.26	7.57	9.60	7.59	9.59	26.52	16.13	7.51	9.54	7.49	9.47	33.35	17.29	7.47	9.43	7.44	9.42
Percentile*	0.000	-	0	0	0	0	0.216	0.055	0.81	0.63	0.65	0.57	0.135	0.007	0.83	0.83	0.66	0.71	0.062	0	0.89	0.85	0.63	0.87
Max	220	10	24	30	24	30	210	67	24	30	24	30	253	85	24	30	24	30	266	102	24	30	24	30
K=3	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	1	-0.04	2	1	2	1	2	2.46	18	24	19	24	3	2.64	15	11	17	11	3	4.82	7	16	9	16
Mean	14.80	1.36	10.83	12.93	13.94	16.36	8.12	7.70	11.31	14.01	13.27	16.21	13.98	12.90	10.96	13.72	14.06	17.19	19.45	19.53	10.47	13.35	14.30	17.74
SD	13.57	0.84	7.24	9.00	7.29	9.33	8.55	4.83	7.33	9.18	7.37	9.23	12.56	6.15	7.41	9.29	7.41	9.26	16.32	8.26	7.41	9.37	7.44	9.33
Percentile*	0.000	-	0.1	0	0.03	0	0.229	0.023	0.74	0.79	0.69	0.71	0.101	0	0.64	0.43	0.56	0.3	0.048	0	0.39	0.58	0.28	0.42
Max	126	5	24	30	24	30	90	38	24	30	24	30	132	52	24	30	24	30	168	64	24	30	24	30
K=4	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	1	-0.04	3	1	3	1	4	1.93	21	26	22	27	2	2.09	11	18	11	19	4	4.09	8	10	9	11
Mean	8.12	0.61	11.25	11.94	13.15	13.97	4.96	4.07	11.96	14.10	13.25	15.50	9.44	8.08	11.38	14.20	13.81	16.84	12.62	12.89	11.14	13.79	14.09	17.11
SD	8.65	0.58	7.45	8.87	7.52	9.44	5.05	2.05	7.37	9.09	7.35	9.22	7.80	3.17	7.35	9.17	7.35	9.18	10.38	4.51	7.44	9.30	7.42	9.35
Percentile*	0.000	-	0.17	0	0.1	0	0.496	0.053	0.83	0.85	0.81	0.83	0.084	0.001	0.48	0.61	0.37	0.54	0.133	0.004	0.39	0.39	0.29	0.3
Max	75	3	24	30	24	30	66	21	24	30	24	30	80	27	24	30	24	30	91	40	24	30	24	30
K=5	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	1	-0.04	4	1	4	1	1	1.73	14	22	14	22	7	1.83	7	10	13	10	4	4.09	5	12	6	13
Mean	3.67	0.20	10.83	11.78	11.54	12.58	3.16	2.73	12.18	14.44	12.93	15.27	6.88	5.88	11.42	14.34	13.29	16.39	9.01	9.70	11.43	13.97	13.73	16.63
SD	6.03	0.40	7.80	9.16	7.95	9.44	3.47	1.20	7.39	9.21	7.39	9.27	5.54	1.97	7.38	9.15	7.37	9.19	7.24	2.97	7.46	9.27	7.47	9.34
Percentile*	0.000	-	0.27	0	0.25	0	0	0.181	0.55	0.72	0.51	0.69	0.597	0.001	0.34	0.37	0.47	0.29	0.213	0.019	0.26	0.44	0.2	0.38
Max	57	2	24	30	24	30	32	11	24	30	24	30	44	17	24	30	24	30	78	26	24	30	24	30
K=6	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Observation	1	-0.04	5	1	5	1	1	0.73	22	4	22	4	1	1.57	6	23	6	23	6	3.64	21	4	22	6
Mean	1.46	0.03	8.60	11.61	8.72	11.75	2.34	1.96	12.28	14.35	12.80	14.88	5.13	4.69	11.75	14.37	13.16	15.92	6.71	7.81	11.51	14.34	13.31	16.36
SD	2.74	0.17	7.97	10.49	8.03	10.54	2.38	1.01	7.38	9.18	7.38	9.24	4.40	1.35	7.39	9.20	7.39	9.24	5.59	2.24	7.47	9.33	7.47	9.35
Percentile*	0.000	-	0.48	0	0.48	0	0	0.037	0.86	0.16	0.83	0.14	0	0.004	0.27	0.75	0.21	0.69	0.496	0.018	0.84	0.18	0.8	0.17
Max	38	1	24	30	24	30	30	8	24	30	24	30	36	13	24	30	24	30	52	21	24	30	24	30

* A percentile position of the observed value in simulated results. Either close to smallest or largest value indicates that the observed value is significant (assume that simulation test has a normally distribution). Note that corners of regions are at (E1, N1), (E2+1, N2+1).

Appendix 5-4 Outputs of observed and simulated maximum subarrays (w = 98 percentile).

	1966						1976						1986						2006					
K=1	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	2	10	13	17	14	19	28	93	13	12	17	19	15	80	13	14	16	19	4	56	14	14	16	16
Mean	1.14	10.07	12.42	15.57	12.49	15.64	5.32	44.82	11.83	14.78	13.23	16.27	2.71	42.77	12.21	15.17	12.86	15.85	1.69	37.66	12.36	15.34	12.65	15.64
SD	0.55	0.35	6.92	8.61	6.91	8.61	5.51	9.43	6.81	8.55	6.79	8.52	2.82	8.81	6.87	8.59	6.86	8.59	1.42	6.53	6.91	8.60	6.91	8.59
Percentile*	0.917	0.000	0.505	0.529	0.502	0.560	0.998	0.999	0.536	0.392	0.591	0.535	0.997	0.997	0.512	0.443	0.566	0.553	0.876	0.963	0.548	0.436	0.575	0.460
Max	8	14	24	30	24	30	36	97	24	30	24	30	22	110	24	30	24	30	12	92	24	30	24	30
Min	1	10	1	1	1	1	1	38	1	1	1	1	1	38	1	1	1	1	1	35	1	1	1	1
K=2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	4	3	14	14	16	16	3	4	16	20	19	21	2	25	16	12	17	14	2	43	13	17	14	19
Mean	1.07	3.00	12.41	15.57	12.44	15.60	1.95	33.34	12.28	15.36	12.62	15.72	1.33	35.48	12.41	15.31	12.55	15.45	1.10	33.21	12.61	15.53	12.66	15.58
SD	0.33	0.24	6.95	8.67	6.95	8.67	2.74	3.90	7.01	8.81	7.03	8.82	1.21	2.03	6.96	8.73	6.96	8.74	0.44	1.25	7.00	8.63	7.01	8.63
Percentile*	0.995	0.024	0.544	0.431	0.587	0.461	0.862	-	0.635	0.632	0.696	0.618	0.871	0.005	0.628	0.375	0.621	0.402	0.935	0.997	0.492	0.532	0.490	0.561
Max	5	5	24	30	24	30	30	55	24	30	24	30	12	53	24	30	24	30	8	52	24	30	24	30
Min	1	1	1	1	1	1	1	8	1	1	1	1	1	8	1	1	1	1	1	19	1	1	1	1
K=3	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	1	2	16	12	17	13	1	2	12	16	13	17	2	9	17	20	19	21	2	27	16	12	17	14
Mean	1.04	1.95	12.45	15.47	12.48	15.50	1.35	23.44	12.50	15.53	12.64	15.69	1.10	30.28	12.56	15.51	12.61	15.55	1.07	29.62	12.46	15.49	12.49	15.53
SD	0.22	0.25	6.96	8.69	6.96	8.69	1.31	6.83	7.03	8.83	7.03	8.82	0.44	7.25	7.00	8.80	7.00	8.81	0.31	4.17	6.97	8.71	6.97	8.71
Percentile*	0.000	0.001	0.624	0.369	0.623	0.369	0.000	0.000	0.456	0.500	0.450	0.494	0.939	0.020	0.659	0.629	0.694	0.627	0.947	0.119	0.627	0.368	0.626	0.398
Max	4	3	24	30	24	30	16	38	24	30	24	30	7	39	24	30	24	30	8	41	24	30	24	30
Min	1	0	1	1	1	1	1	2	1	1	1	1	1	8	1	1	1	1	1	8	1	1	1	1
K=4	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	1	1	11	12	12	13	1	2	20	22	21	23	2	2	12	15	13	17	2	8	17	20	19	21
Mean	1.03	0.92	12.33	14.39	12.35	14.40	1.20	12.06	12.30	15.25	12.39	15.35	1.15	15.02	12.37	14.20	12.44	14.27	1.09	17.54	12.40	15.48	12.44	15.52
SD	0.17	0.27	6.96	8.94	6.96	8.94	0.71	4.88	6.96	8.77	6.96	8.77	0.49	4.77	6.95	8.58	6.95	8.59	0.34	3.70	6.88	8.69	6.88	8.69
Percentile*	0.000	0.000	0.430	0.416	0.430	0.415	0.000	0.000	0.796	0.703	0.792	0.699	0.894	0.000	0.460	0.527	0.456	0.557	0.931	0.000	0.673	0.633	0.713	0.631
Max	3	1	24	30	24	30	10	30	24	30	24	30	7	34	24	30	24	30	4	29	24	30	24	30
Min	1	0	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1
K=5	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	1	0	12	13	13	14	1	1	17	15	18	16	1	2	14	19	15	20	1	2	12	15	13	16
Mean	1.00	0.00	11.74	3.03	11.74	3.03	1.10	6.62	12.44	13.99	12.49	14.03	1.12	8.07	12.41	13.76	12.47	13.82	1.08	9.67	12.44	15.37	12.48	15.41
SD	0.00	0.00	6.96	2.27	6.96	2.27	0.43	2.18	6.94	8.84	6.95	8.84	0.38	0.94	6.96	8.49	6.96	8.50	0.30	1.31	6.91	8.67	6.91	8.67
Percentile*	0.000	0.000	0.506	0.996	0.506	0.996	0.000	-	0.670	0.540	0.667	0.538	0.000	0.000	0.546	0.685	0.544	0.683	0.000	0.000	0.456	0.474	0.454	0.472
Max	1	0	24	16	24	16	6	15	24	30	24	30	6	16	24	30	24	30	4	19	24	30	24	30
Min	1	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K=6	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2	A	S	E_1	N_1	E_2	N_2
Obs.	1	0	16	13	16	14	1	0	18	15	19	16	2	1	17	15	19	16	1	1	14	13	15	14
Mean	1.00	0.00	12.53	5.35	12.53	5.35	1.07	3.76	12.47	10.17	12.50	10.20	1.04	7.19	12.55	18.03	12.57	18.05	1.05	7.62	12.51	15.34	12.53	15.37
SD	0.00	0.00	6.93	3.01	6.93	3.01	0.28	2.36	6.95	7.91	6.95	7.93	0.21	1.62	6.90	8.27	6.90	8.27	0.24	1.15	6.90	8.70	6.90	8.70
Percentile*	0.000	0.000	0.623	0.974	0.623	0.974	0.000	-	0.709	0.733	0.708	0.731	0.961	0.000	0.667	0.346	0.705	0.346	0.000	0.000	0.541	0.409	0.540	0.407
Max	1	0	24	21	24	21	5	9	24	30	24	30	4	10	24	30	24	30	4	10	24	30	24	30
Min	1	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

* Percentile position of the observed value in simulated results. If close to the smallest or largest value, indicates that the observed value is significant (assuming that the simulation test has a normal distribution). Note that corners of regions are at (E1, N1), (E2+1, N2+1).



Chapter 6. Singular Spectrum Analysis for decision tree classification

This chapter introduces the use of the unique statistical decomposition method, Singular Spectrum Analysis (SSA), as a pre-processing method for noisy input data to improve data mining classification. Two distinctive environmental science problems are demonstrated. The first study introduces how pre-processing a noisy climate time series input via SSA can improve prediction of air pollution levels by a decision tree algorithm. The second study introduces how noisy imagery data, constructed from red, green and blue histograms extracted from image tiles, can be processed by SSA then used by a decision tree classifier to predict areas of defoliation caused by the mountain pine beetle in aerial forest imagery. Results of predicting the defoliated areas are then compared with a pixel-based clustering method. Both studies suggest that SSA inputs produced improved predictions. The use of such decomposition methods will help computer algorithm prediction in other applications.

Study I. Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution (Fukuda 2007)

6.1. Introduction

This study covers the brief concept of the mathematical decomposition method, Singular Spectrum Analysis (SSA), as a data pre-processing method to help improve the C4.5 algorithm to predict carbon monoxide (CO) levels using various noisy climate measurements. The constructed decision tree was then investigated to discover knowledge about the relationship of climate and CO levels.

As previously discussed in Chapter 4, the causal effect of air pollution on human and environment health is a worldwide problem. Air pollutant levels even below standard concentrations are known to affect human health, with increases in respiratory symptoms, chronic cough, bronchitis and chest illness, and deterioration in pulmonary function (Koening 2000). This chapter investigates how pre-processing various noisy climate attributes via the unique decomposition method, Singular Spectrum Analysis (SSA), improves the ability of a decision tree classifier algorithm to predict levels of air pollutant carbon monoxide (CO). Chapter 4 investigated the maximum association of air pollution, particulate matter (PM) and sulfur dioxide (SO_2), various climate attributes and health (respiratory admission rate) using the *K*-Maximum Subarray Algorithm (*K*-MSA).

The study area, Christchurch, New Zealand, with a population of about 334,000 and an area of 452 km^2, suffers from a serious winter air pollution problem due to domestic heating, e.g., burning wood and coal, and poor air dispersion due to a combination of winter weather and its topographic factors, primarily a medium sized hill located adjacent to the city, which traps air pollutants in a temperature inversion layer; see details in Kossmann and Sturman (2004) and Chapter 4. The main winter air pollutants are CO from domestic heating and motor vehicles, PM from domestic heating, SO_2 from industry and NO_2 (a product of the oxidation reaction of NO) from motor vehicles (Scott and Gunatilake 2004). Recent investigation of particulate matter of diameter below 10 μgm^{-3} (PM_{10}) and the acute respiratory morbidity rate in the study area reports that even low PM_{10} levels (less than 10 $\mu g/m^3$) can impact on different age ranges, in particular, very young (under five years) and older ages (55 years and over), with an association that varies between female and male and by season (Fukuda and Takaoka 2007). Also, time series analysis using SSA shows that short and long-term air pollution levels are affected by changes in both local climate and global climate (Fukuda 2004; Fukuda and Hudson 2005).

In recent years, data mining, a process of knowledge discovery in databases (KDD), is also found to be a useful tool among environmental scientists (Fukuda and Pearson 2006a,b; Spate et al. 2006) due to its flexibility to handle problems in environmental systems, which are often ill-structured and non-linear domains (Spate et al. 2006), and involve multidisciplinary factors, e.g., global and local ecological, social and economical factors. To investigate the air pollution and climate data set that is generally noisy and skewed, a primary step is to reduce the noise, although determining the noise component of such a noisy and skewed structure can be difficult. Attribute selection can be used to remove the outliers as a data pre-processing step, but it may lose the time sequence, as the air pollution and climate time series are associated, day-to-day. Hence, smoothing methods are commonly applied to climate and air pollution studies. One of the common smoothing techniques is Generalized Additive Models (GAMs) (Aldrin and Haff 2005), which is a statistical method for smoothing non-linear time series, and is used to identify response-predictor relationships. Recently, Li and Shue (2004) used the wavelet transform as a data pre-processing step to extract the trends of air pollution levels in order to apply further neural network models, since data mining algorithms with pre-processed data sets generally work efficiently to provide improved results (Li and Shue 2004, Li et al. 2002).

In this study, two investigations are carried out. Firstly, Singular Spectrum Analysis (SSA) is introduced as the noise reduction approach to pre-process data prior to applying a data mining technique, the C4.5 decision tree classifier (Quinlan 1993). The noisy climate time series is decomposed and separated out from noise by SSA to form several additive components, which are used to construct decision trees to predict the different air pollution levels of carbon monoxide (CO). Decision tree classification accuracy is then examined to see how SSA helped the algorithm. Secondly, the obtained decision trees are examined to provide threshold climate values that impact on different CO levels. The investigation helps support knowledge on the cause and effect relationship of climate and air pollution profile.

6.1.1. Singular Spectrum Analysis for data mining input

Singular Spectrum Analysis (SSA) is an innovative model-free nonparametric method of time series analysis, a mixture of mathematical and statistical analyses: namely classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing (Golyandina et al. 2001). For example, SSA has been traditionally applied to digital signal processing (Kumaresan and Tufts 1980) and oceanographic research (Colebrook 1978). Recently it has been applied to an air pollution study (Fukuda 2004; Fukuda and Hudson 2005b), and SSA decomposed structures have been applied to data

mining techniques for image segmentation (Fukuda and Pearson 2006a,b). However, in this study, SSA is used for data pre-processing to help the data mining algorithm by removing noisy structures from the data set.

SSA provides two benefits; the decomposed structures help improve the results of the decision tree algorithm and SSA helps identify noise in the structures, because it decomposes the noisy time series into several *additive* components – separating out several high and low frequency signals from the original time series – that can be grouped and are reconstructed to form the new time series. Note that the signals obtained by SSA decomposition differ from those obtained by filtering out frequency bands with the Fourier transform, as they are generated from eigenvectors and as such are not purely related to frequency. This facilitates data exploration by adding or removing such *additive components* (low to high frequencies) to construct the decision tree. During this process, it identifies *which* components can potentially be noise, and the improvement can be examined by the classification accuracy. For example, adding insignificant components (generally high frequency) to the main structures (low frequency) can lower or have no influence on the classification accuracy. On the other hand, removing significant components (including some high frequencies) may lower the classification accuracy, which suggests that these components are unlikely to be noise.

6.1.2. Knowledge discovery for climate and air pollution

Extracted decision trees with high classification accuracies are investigated to understand the cause and effect relationship between climate and air pollution levels. This is carried out by examining *how* the decision pathway of climate attributes contributes to change in air pollution levels, such as *which* climate attributes influence air pollution levels, to *what* degree. Note that this study aims to provide knowledge from examining the decision trees via a data mining tool rather than providing prediction rules, to be used to predict air pollution levels in an unknown data set. This is because the studied data set is not large enough to demonstrate accurate prediction rules, but it can be at least used as a knowledge discovery tool.

To enhance the relationship between climate and air pollution level, decision trees are generated from the training data sets of SSA components that are each made up of a single season (dividing the annual data set into four seasons) and the annual data set (all seasons), to compare how the decision pathways of climate influence air pollution levels differently as well as differences in the classification accuracy among different seasons.

6.2. Data and methods

The following section will discuss the studied data, the concept of the Singular Spectrum Analysis (SSA) in brief, the climate attributes decomposed by SSA for use as input attributes for the decision tree classifier to predict CO level, and the method for assessing the obtained decision tree structures.

6.2.1. Studied data

Four years (October 1998 – September 2002) of air pollution and climate daily measurements were provided by an Environment Canterbury (ECan) air pollution monitoring station, located in a residential area, Coles Place, in Christchurch. Six climate measurements are used as input attributes to predict the CO levels: relative humidity (RH in %), temperature measured at 1m above the ground (TG in C°) and at 10m above the ground (TT), the temperature difference (TD = TG-TT), wind speed (WS in m/s), and wind direction (Wdir, measured in degrees: 0° and 360° for north, 90° for east, 180° for south, 270° for west). The original climate time series and the original CO time series are shown in Fig. 6-1, left and right, respectively. Fig. 6-1, left shows six climate attributes from left to right, RH to Wdir along the *x*-axis, where each climate attribute time series covers, from left to right, October 1998 to September 2002. Negative values of TD (Fig. 6-1, left) indicate the formation of a temperature inversion, which traps air pollutants under a layer of warmer air. The CO levels are categorised into three levels based on the lower and upper quartile (LQ and UQ), since its distribution is rightward skewed; low (L) ≤ 0.14 mg/m^3 at LQ, medium (M) ≤ 0.70 mg/m^3, and high (H) > 0.70 mg/m^3. Generally, all six time series were noisy, and CO, TG, TT and TD show reasonably strong seasonal structures with some high frequencies (Fig. 6-1, left). Note that all data were scaled by dividing each value by the maximum in order to improve the ease of comparison between the SSA results of the climate and air pollution data.

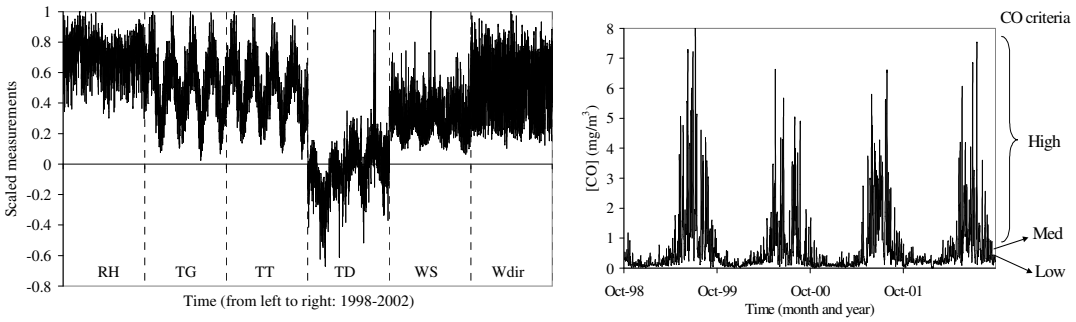


Fig. 6-1 Six different climate time series (left) and original time series of CO (left).

Six climate measurements are listed from relative humidity (RH), temperature above 1m (TG) and 10m (TT) above the ground level, temperature difference (TG-TT), wind speed (WS) and wind direction (Wdir).

6.2.2. Singular Spectrum Analysis

SSA was used to decompose the six climate time series to create input data sets (several additive components) for further data mining application. The SSA procedure has four steps (Golyandina et al. 2001; Fukuda 2004). The first step is embedding, which transforms the original one dimensional time series,

$$F = (f_i) = (f_1, \ldots, f_N) \tag{6-1}$$

into an L-dimensional series,

$$X_i = (f_{i-1}, \ldots, f_{i+L-2})^T, \tag{6-2}$$

where $1 \le i \le K = N - L + 1$ and L is the window length ($\le N/2$). The embedding process turns the one-dimensional time series F into the L-trajectory matrix,

$$X = [X_1 : \ldots : X_K] \tag{6-3}$$

which can be rewritten as,

$$X = (x_{ij})_{i,j+1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix}. \tag{6-4}$$

Note that the matrix X is a Hankel matrix, which has equal elements on the diagonals ($i+j = const.$).

The second step of SSA is to decompose the obtained trajectory matrix by the singular value decomposition (SVD). Let $U_1, \ldots, U_L$ represent the corresponding orthogonal eigenvectors of the matrix $S=XX^T$. Then denote V_i as the eigenvector of S, which corresponds to the eigenvalue λ_i for $i = 1, ..., d$, where d is the number of nonzero eigenvalues ($d < L$ and $1 \le i \le d$),

$$V_i = \frac{1}{\sqrt{\lambda_i}} X^T U_i, \tag{6-5}$$

then the result of the *SVD* of the trajectory matrix, X, becomes

$$X = X_1 + \cdots + X_d, \tag{6-6}$$

where

$$X_i = \sqrt{\lambda_i} U_i V_i^T. \tag{6-7}$$

The i^{th} eigentriple (ET) is constructed from the three attributes in equation 6-7 that make up X_i: the singular value (the square root of the i^{th} eigenvalue) and two orthogonal vectors, the i^{th} right (V_i) and left (U_i) singular vector of the trajectory matrix. Note that each ET has a

different variance, and the sum of the variances for all ETs is 1. These are the additive components.

In the third step, similar ETs are grouped together. It is important to combine appropriate ETs, or in other words, keep the components as similar as possible, rather than mixing dissimilar components, e.g., mixing low and high frequencies, because it decreases the quality of the results in reconstructing the new time series (step 4). In this study, the ET grouping procedure was performed computationally using FastGrouping, a separately developed program that uses Fourier expansion to determine ET similarity (Fukuda 2004). The Fourier expansion (Golyandina et al. 2001, Fukuda 2004) provides a correlation coefficient, $\rho_{1,2}$, which is calculated from the cross power of the two series, $F = F_{(1)}+F_{(2)}$, obtained from different ETs in equation 6-8;

$$\rho_{1,2} = \sum_{k=0}^{N} \sqrt{\Pi_{f_1}^{N}(k/N)}\sqrt{\Pi_{f_2}^{N}(k/N)} \le \left\|F^{(1)}\right\|\left\|F^{(2)}\right\|. \tag{6-8}$$

The normalized form of equation 6-8 is

$$\rho R_{1,2} = \frac{\sum_{k=0}^{N} \sqrt{\Pi_{f_1}^{N}(k/N)}\sqrt{\Pi_{f_2}^{N}(k/N)}}{\|f_1\|\|f_2\|}, \tag{6-9}$$

and the magnitude of $\rho R_{1,2}$ indicates the similarity of the spectra of the relevant two signals. Each eigentriple is successively paired with every other eigentriple, and for each pair of eigentriples, the value of $\rho R_{1,2}$ is computed (in equation 6-9) using the pair of eigenfunctions (eigenvalues and eigenvectors) as $F_{(1)}$ and $F_{(2)}$. It is then computed again using the pair of principal components. Averaging the resulting two $\rho R_{1,2}$ values provides a single metric, which improves the sensitivity. This provides more reliable results than when the eigenfunctions and the principal components are considered separately. Next, the $\rho R_{1,2}$ value is compared with a threshold (between 0.50 and 0.90) and the two eigentriples are placed in the same group if the metric is greater than the threshold. Lowering the threshold provides fewer ET groups, grouping ETs less accurately, and raising the threshold gives the opposite. Generally, a threshold between 0.70 and 0.85 is recommended (Fukuda, 2004).

The fourth and final step is called diagonal averaging. It is a linear operation for reconstructing time series from the additive components and ET groups that are chosen in step 3,

$$F = F_1 + \ldots F_m, \qquad 1 \le i \le \mathrm{m}. \tag{6-10}$$

Each of the six climate time series is decomposed by SSA into a number of additive components (each constructed from a single ET or a group of ETs, and of the same length as the original time series, *F*), which are used to generate decision trees as follows.

6.2.3. SSA for the decision tree classifier

To investigate the effectiveness of using SSA for data pre-processing for data mining, a decision tree classifier was applied on climate attributes to predict three CO levels (high, H; medium, M; and low, L), and the classification accuracy was used to assess the improvement. Results were compared for the original time series (without the SSA data processing) and the SSA additive component time series. From each time series, the full length of the time series was divided into four seasons to compare the annual data set (full data set) and seasonal data sets. Hence, the following procedure, generating a decision tree, was repeated for a total of five data sets (one covering the whole year, and one for each of spring, summer, autumn and winter), for the original and each of the additive component time series.

Each data set was divided into three parts, and three training and three test data sets were created. For example, the first training data set consisted of the first two thirds of the data set, and the first test set consisted of the remaining third. Thus, three distinct training and test data sets were created. A decision tree classifier, J4.8 from WEKA (Witten and Frank 2005), based on the C.4.5 algorithm (Quinlan 1993), was used to generate a decision tree from each training data set, and was tested on the test data set to provide a classification accuracy. The average and standard deviation (SD) of the three classification accuracies obtained from three test data sets were used for the results.

The specific procedure (repeated for each training set) for generating decision trees for experimenting with the noise reduction method via additive components was as follows. Firstly, a decision tree is generated from a single data set, which covers a full year or a single season. Secondly, decision trees are generated from a data set for each additive component, first removing the structures for ET151-180 from the rest (ET1-150), and increasing the range of eigentriples removed until reaching ET3-180, leaving only ET1 and ET2 (ET1 is kept to provide a base for the components). Hence, six experiments are repeated to generate six single decision trees for each of the five data sets (the full and seasonally divided data set). Note that the experiment starts from removing the 8th additive component (ET151-180 in Fig. 6-2, H), and the 1st additive component (ET1 in Fig. 6-2, A) is not removed.

6.2.4. Knowledge discovery from decision trees

To introduce the outcome of applying the data mining technique on climate and air pollution, the decision trees were examined in detail to increase understanding about the cause and effect relationship of climate and CO levels. Decision pathways, such as which climate attribute is most responsible for the high CO level, can be investigated. Note that investigations in this study are carried out by examining the decision tree with the best classification accuracy out of each group of three training data sets. Also to simplify results, the decision pathway was focused and summarised on only the *high* CO level, while the decision trees classify CO into three levels (H, M and L); results on M and L are not described here. To contrast seasonal climate impacts on the high CO level, examination of decision trees is focused on seasonally divided data sets, thus the full data set is not interpreted.

6.3. Results and discussions

6.3.1. Extraction of additive components

In this study, a window length, *L*, of 30 (~one month) was selected, because it was one of the dominant frequencies of the air pollution time series. FastGrouping with a threshold of 0.85 provided a number of ET groups. Eight heterogeneous ET groups (and variances, shown as percentages, in brackets) were extracted as input data sets for the data mining application, shown in Fig. 6-2, A to H respectively: ET1 (90.2%), ET2 (2.72%), ET3 (0.68%), ET4-40 (4.14%), ET41-80 (1.46%), ET81-126 (0.70%), ET127-150 (0.11%), and ET151-180 (< 0.01%).

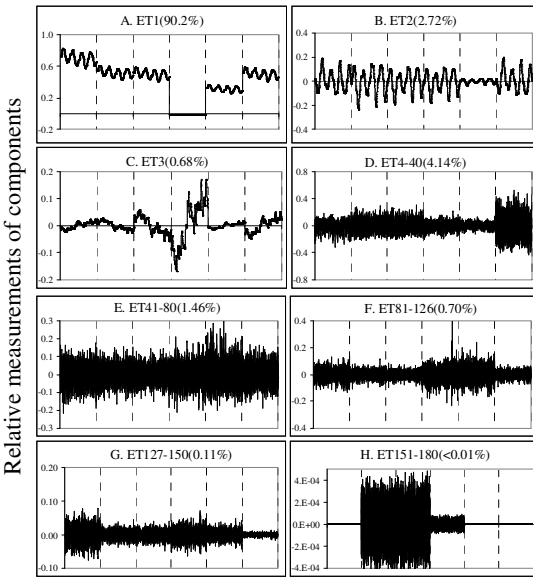


Fig. 6-2 Eight SSA climate additive components made by ET groups.

Note that the dotted lines indicate a single climate attribute, RH, TG, TT, TD, WS and Wdir from left to right. The variance of each ET group is shown in brackets.

Fig. 6-2 shows six climate attribute ETs of RH to Wdir in the same manner as Fig. 6-1, left. Generally, the first three additive components (ET1 to ET3) hold important structures. The first eigentriple, ET1, which is made by the lowest frequency, has a large variance, describes the general trends, and provides the base structure, so it is always added to the other ETs. ET2 and ET3 generally describe the seasonal structure and change points, or structural changes respectively (Fukuda 2004). The ETs after ET4 are made by high frequencies with reasonably small variances,

thus they are grouped with similar components to form larger components. Note that these structures are generally not used, if the purpose of the study is to extract the smooth time series (see details in Fukuda 2004).

6.3.2. Comparison of classification accuracies

Table 6-1 shows summary results of decision tree classification accuracy (in %) using the original climate time series and different grouped additive components (ETs) based on the full and seasonally divided data sets.

Successively larger numbers of ET groups are removed from Case 1 (removing < 0.01% of the entire structure) to Case 6 (removing 7.10% of the entire structure) in Table 6-1. Table 6-1 also shows the proportion (in %) of each CO level; H, M, and L, as a brief indicator. Generally, application of the decision tree classifier is better than simply guessing the CO levels, if its score is better than this number. Table 6-2 shows the confusion matrix for the best classification accuracy within the full data set and each seasonally divided set. Note that the number of instances is the sum of three test data sets.

6.3.3. Full length and seasonally divided data.

Dividing the full data set (annual) into seasons shows the different classification accuracies among seasons. The mean and standard deviation (SD) of the original and all cases in Table 6-1 (bottom rows) show that the highest classification accuracy is found from winter (76.4±5.7%), which is higher than applying the full length of the data (66.7±2.7%). In fact, the average classification accuracy of the seasonally divided data (67.8±2.7%) is found to be slightly higher than simply applying the full data set, shown in Table 6-1 (bottom rows). This suggests that even though the sample size has became one fourth of the full length of the data, separating out seasons in environmental data sets that have seasonality helps the algorithm by highlighting relevant characteristics. On the other hand, the lowest classification accuracies compared with the full data set are found from spring (61.7±3.1%) and summer (63.5±8.0). This may due to the low proportion of high CO levels (9.3% and 0.6%) in the spring and summer data, although the reason for the higher classification accuracy in summer compared to spring may be that the summer data set consists of almost half M (44.9%) and half L (54.6%), as it only predicts either M or L, which may confuse the algorithm less, compared with the spring data set.

Table 6-1 Summary of decision tree classification accuracy (CA) using different SSA decomposed components.

(%)	Spring	Summer	Autumn	Winter	Mean (all seasons)	Full data set
Original proportion of CO levels						
H	9.3	0.6	28.5	60.9	24.8	25.0
M	59.3	44.9	58.7	36.4	49.8	49.8
L	31.3	54.6	12.8	2.7	25.3	25.2
Original time series						
CA	61.5	60.7	71.2	80.1	68.4	67.8
SD	5.0	4.7	3.7	3.3	1.0	7.9
Case 1. Removing ET151-180 (<0.01%) in Fig. 6-2, H from the rest (= adding G. ET127-150 on ET1-80)						
CA	61.5	60.7	71.2	80.1	68.4	67.8
SD	5.0	4.7	3.7	3.3	1.0	7.9
Case 2. Removing ET127-180 (0.12%) in Fig. 6-2, G and H from the rest (=adding F. ET81-126 on ET1-80)						
CA	62.4	59.5	71.2	83.4	69.1	66.9
SD	4.1	4.5	1.6	1.3	2.5	9.3
Case 3. Removing ET81-180 (0.82%) in Fig. 6-2, F to H from the rest (=adding E. ET41-80 on ET1-40)						
CA	60.7	55.7	69.0	79.3	66.2	65.9
SD	1.2	6.5	7.2	3.0	2.3	8.9
Case 4. Removing ET41-180 (2.28%) in Fig. 6-2, E to H from the rest (=adding D. ET4-40 on ET1-3)						
CA	56.0	58.5	67.1	68.2	62.5	62.0
SD	2.8	1.7	2.2	4.2	2.7	5.3
Case 5. Removing ET4-180 (6.42%) in Fig. 6-2, D to H from the rest (= adding C. ET3 on ET1-2)						
CA	63.5	77.3	68.7	72.8	70.6	70.8
SD	1.6	0.5	0.6	2.2	0.5	5.1
Case 6. Removing ET3-180 (7.10%) in Fig. 6-2, C to H from the rest (= adding B. ET2 on ET1)						
CA	66.2	72.0	68.5	70.9	69.4	65.8
SD	2.5	1.0	1.1	2.0	1.6	2.2
Mean: the original and all cases within the same season	61.7	63.5	69.6	76.4	67.8	66.7
SD: the original and all cases within the same season	3.1	8.0	1.6	5.7	2.7	2.7

Table 6-2 Comparison of the confusion matrices between the original time series and the high frequency separated SSA additive components for all data sets.

Note that the total number (sum of all three test results) of instances is shown. Numbers in bold indicate correctly classified instances.

Spring				Summer				Autumn				Winter				Full data set			
Original time series																			
	H	M	L		H	M	L		H	M	L		H	M	L		H	M	L
H	**12**	11	1	H	**1**	0	0	H	**66**	15	0	H	**189**	26	0	H	**276**	67	2
M	22	**143**	44	M	1	**67**	46	M	39	**185**	36	M	35	**106**	10	M	86	**498**	150
L	0	62	**69**	L	0	95	**151**	L	0	16	**11**	L	0	2	**0**	L	3	163	**216**
Case 6 (Remov.ET3-180)				Case 5 (Remov.ET4-180)				Case 2 (Remov.ET127-180)				Case 2 (Remov.ET127-180)				Case 5 (Remov.ET4-180)			
	H	M	L		H	M	L		H	M	L		H	M	L		H	M	L
H	**10**	4	2	H	**0**	1	0	H	**72**	24	0	H	**187**	13	0	H	**261**	102	9
M	21	**154**	35	M	1	**98**	16	M	33	**177**	34	M	37	**120**	10	M	99	**539**	124
L	3	58	**77**	L	1	63	**181**	L	0	15	**13**	L	0	1	**0**	L	5	87	**235**

6.3.4. Removal of noisy components for the CO prediction

For generating decision trees, removing some proportion of the structures (from < 0.01% in Table 6-1, Case 1 to up to 7.10% in Table 6-1, Case 6) from the original time series has shown some improvement over the original time series, although it varies by seasons and different additive ETs (Table 6-1). For each series, the classification accuracy peaks after a certain number of high frequencies have been removed.

This point can be used to identify which structures are significant to capture the best decision trees. The original times series and Table 6-1, Case 1 show the same classification accuracy for all data sets. Removing smaller high frequency structures, ET151-180 (< 0.01% in Fig. 6-2, H), would not influence the classification accuracy, and may not help the algorithm. This suggests that ET151-180 structures may be noise, or insignificant.

- **Prediction of summer CO**

Table 6-1, Case 5, the summer data, shows the most significant classification accuracy improvement. Removing ET4-180, a total of 6.42% of the structure (in Fig. 6-2, C to H) shows the accuracy is 77.3%, up to a 16.6% improvement compared with the original time series classification accuracy (60.7% in Table 6-1). As previously mentioned, the average summer classification accuracy was the lowest. However, the summer data set is made up almost completely of two CO levels, M and L, and it contains fewer outliers and high pollution levels compared with winter. Hence, removing most of the high frequencies (including potential noise) or outliers that are obtained from 6.42% of the structures, ET4-180, may help the algorithm. Most of the classification errors are between M and L; 63 instances of L and 16 instances of M were misclassified as M and L respectively, but these errors are greatly reduced compared to the original data set, where 95 and 46 instances of L and M were misclassified respectively (see Table 6-2; original and Case 5, summer).

- **Prediction of spring CO**

The spring data set (Table 6-1, Case 6) also shows similar findings, but the classification accuracy is lower (66.2%) than summer, and the improvement was 4.7% compared with the original time series (61.5%). However, the improvement for the spring data set is obtained from removing 7.10% of the structures, ET3-180. Note that the spring data set contained about 9.3% high CO levels (Table 6-1), although the algorithm works better with dominant low frequencies, ET1 and ET2 (Fig. 6-2, A and B), describing the seasonal oscillation by removing most of the high frequencies. However, interestingly, the use of only low frequencies improves the misclassification between M and H; 4 instances of H were misclassified as M, compared to 11 for the original time series (Table 6-1; original and Case 6, spring).

- **Prediction of winter CO**

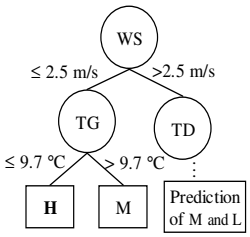
The winter (Table 6-1, Case 2) and full data set (Table 6-1, Case 5) classification accuracies (83.4% and 70.8% respectively) show about 3% improvement by removing 0.12% of high frequencies, ET127-180, and removing 6.42% of structures, ET4-180, compared with the original time series (80.1% for winter and 67.8% for the full data set). While the winter data set contains many high CO data points (60.9%), removing further high frequencies that are obtained after ET81 (Table 6-1, Case 3) decreases the classification accuracy, as it may remove truly high CO levels, which should not be considered as outliers. However, an interesting point is that removing high frequency eigentriples with very small variance, e.g. ET127-180, with 0.12% (Table 6-1, Case 2) shows an improvement, increasing the classification accuracy by about 3%. This may suggest that ET127-150 may be potential noise. From this improvement, the correct classification for M is increased from 106 to 120 instances (Table 6-2; original and Case 2 in winter). However, no correct classification for L is observed, which needs further investigation. The full data set shows higher classification accuracy as more high frequencies are removed up to ET4-180. Since the full data set lacks characteristics compared with seasonally divided data sets, removing all frequencies except the general trend (ET1), seasonal components (ET2) and change points (ET3) provides smoothed but detailed time series structures that help to generate the decision tree with the best classification accuracy. The major classification improvement resulted from increasing the number of correctly classified M instances from 498 (original series) to 539 by decreasing misclassification between L and M (Table 6-2; original and Case 5).

- **Prediction of autumn CO**

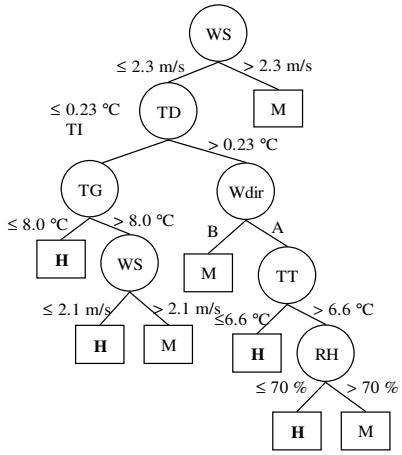
An interesting observation is seen from the autumn data set. Removing any components did not change the classification accuracy, although removing 0.12% of ET127-180 (Fig. 6-2, G and H) kept the same classification accuracy as the original time series (71.2%). Therefore, these structures could be considered as potentially insignificant noise that can be eliminated even without changing the structures, and removal of these may or may not help the algorithm, because the variance of these structures are very small (0.12%). However, the differences between the original and removing ET127-180 (Table 6-2, Case 2) is that removing ET127-180 improves detection of H, increasing correctly classified instances from 66 to 72, but it decreases the correctly classified instances of M from 185 to 177. This point needs further investigation (Table 6-2, Case 2 for winter).

Overall, removing more high frequencies from the original time series improves the classification accuracy for spring, summer and the full data set. In particular, spring and

A. Spring data set.



B. Autumn data set.



C. Winter data set.

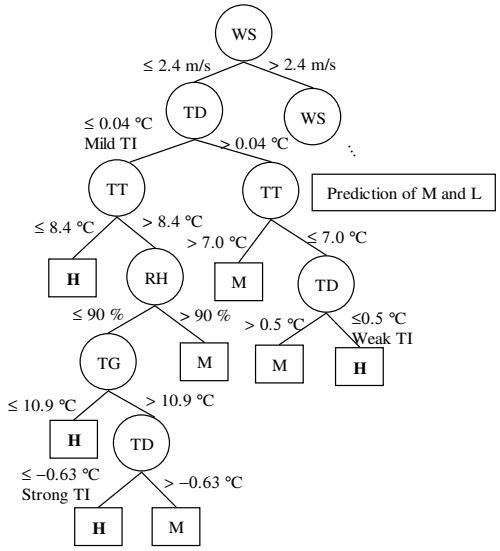


Fig. 6-3 Examples of decision trees for spring (A, top), autumn (B, middle) and winter (C, bottom).

Note that all decision trees shown here focused on the high CO level, otherwise branches for predicting medium and low CO levels are not shown. The autumn decision tree (B) shows the wind direction as A (NE, E, SE direction) and B (S, SW, W, NW direction).

summer contain fewer high levels of CO, so removal of high frequencies such as potential noise, outliers or insignificant signatures helps the algorithm efficiently. For example, the maximum classification accuracy improvement was 16.6% for summer by removing 6.42% of the structures (most of the high frequencies), compared to non pre-processed data. On the other hand, removing only a very small amount of high frequency information, the ET127-180 structures (0.12%), improved the winter classification accuracy by 3%, compared to non pre-processed data. The autumn data set did not show any particular improvement, and it may require further investigation. The use of SSA additive components as inputs for generating decision trees may have future use for any noisy time series, as this provides better classification accuracy for some parts of the data, which can be helpful for the overall analysis. It also allows exploring the data set.

6.3.5. Knowledge discovery from decision trees

Fig. 6-3 shows the highest performing decision tree out of the three training data sets for each of spring, autumn and winter (with respective accuracies of 68.6%, 73.0% and 84.6%). Note that only partial decision tree structures are shown in Fig. 6-3, constructed by the branches that predict high CO levels using various climate variables. Therefore, the following interpretations describe the relationship between climate and high CO level. Also, note that the following results do

not discuss results from summer, because the summer data almost entirely consisted of M and L CO levels.

Since the input climate attributes are numerical, the decision trees have numerical threshold values. Note that the autumn decision tree (Fig. 6-3, B) shows wind direction (Wdir), which takes the value of A for easterly and southeasterly wind and B for southerly, southwesterly and westerly wind. The dominant wind direction is southeasterly, followed by southerly, southwesterly, easterly, and westerly. As previously mentioned in Section III, negative values of TD indicate the formation of temperature inversion (TI). The winter decision tree (Fig. 6-3, C) shows three TD nodes; the lowest (most negative) TD value suggests a strong TI (≤ -0.63 °C), whereas smaller (≤ 0.04 °C) and larger (≤ 0.5 °C) positive TD values suggest the mild and weaker TI formation.

The winter decision tree has the largest tree size (*TS*=29) of all the trees and the highest number of leaves (*NL*=15), suggesting that the decision process for the winter CO level is most complicated, whereas the spring decision tree has the simplest and smallest tree (*TS*=11 and *NL*=6), and the autumn decision tree (*TS*=15 and *NL*=8) lies between the spring and winter trees.

Common climate responses to the high CO level are found. The most important climate factor (found at the root of the tree) is WS with the value of ≤ 2.3-2.5 m/s (the threshold varies between seasons). The mean and standard deviation of wind speed in the study area are 2.60±0.97. Hence, when the wind speed is lower than the mean (light wind speed) the CO level is high. The second most important climate attribute is TD. The autumn data set has milder TI formation (≤ 0.23 °C) than winter, as generally TI is often observed more in winter with lower temperatures. Three different levels of TI (strong, medium and weak) also associate with the high CO level. However, the spring decision tree shows the association of TD is more with M and L (Fig. 6-3, A). In spring, the ≤ 9.7 °C TG is responsible for the high CO level instead. Interestingly, only the autumn decision tree uses the wind direction attribute; southeasterly wind associates with high CO level via lower TT (≤ 6.6 °C), but when TT is above 6.6 °C with lower humidity (≤ 70%; dryer air), the association of the high level is detected (Fig. 6-3, B). A similar finding is found from winter (Fig. 6-3, C). The association of the high CO level is: during mild TI, via lower TT (≤ 8.4 °C); during dryer relative humidity (≤ 90%), via colder TD (≤ 10.9 °C) or via further strong formation of TI (≤ -0.63 °C). Also the weaker TI associates with the high CO level via lower TT (≤ 7.0 °C).

Overall, the climate attributes responsible for the CO level are light wind speed and temperature inversion formation. This is a reasonable finding, also seen from previous

research in the study area (Fukuda 2004). As this study is the first attempt for applying the data mining technique, decision trees for knowledge discovery on the climate and air pollution, it is important to note that the exact threshold values and findings require further investigation, carried out by experts in this field.

6.4. Conclusions

The use of SSA as the noise reduction method for the data mining application, a decision tree classifier, successfully improved the classification accuracy in this climate time series, compared with the original time series. The improvements were more effective when the data set (containing all four seasons) was divided into seasons. The summer data set classification accuracy improved up to 16.7%, compared with the original time series, after removing 6.42% of the signal. However, the autumn data did not show any improvement, which may suggest that other attributes can describe the CO level better than the currently used climate attributes. The advantage of using SSA is to provide several additive components that can be added to or removed from the main structures, allowing exploration of the nature of the noisy time series data set.

Observing how the classification accuracy changes provides information on which components are essential to generate the decision tree or alternatively, which components are insignificant signatures in the noisy time series (potential noise). In this application, generating the decision trees using climate attributes to predict the CO levels from different seasons provided knowledge of the responsible climate attributes or the pathway for the CO levels. In particular, the decision tree provides threshold values of each climate attribute that are responsible for the change of CO levels. Detailed examination of the decision trees suggests that the most important climate condition is wind speed less than or equal to 2.3 to 2.5 m/s, which associates with high CO levels. The second most important climate attribute is any level of temperature inversion formation. Note that the exact threshold value for each climate attribute requires further investigation from experts in the field, although results may be useful as indexes for future climate and air pollution study. In order to increase the sensitivity in generating the decision tree, the fuzzy decision tree technique may help reduce misclassification of the different CO levels (H, M, and L). However, the introduced noise reduction method via SSA is an encouraging data pre-processing method for any data mining techniques. The data mining approach in this study can be adapted and used as a knowledge discovery tool for various environmental researches in future. A new hybrid prediction model will be developed in the near future to incorporate such mathematical and statistical methods,

and computer algorithms, to investigate air pollution, climate and health. This will be discussed in Chapter 7.

6.5. Acknowledgment

Thanks to P. Pearson for giving me useful comments as well as helping with programming for the data preparation and editorial work. Thanks to Dr. B. Martin, and Chloe and Jack Duckling for their inspiration and encouragement towards this study. Thanks to T. Aberkane (ECan), Dr. A McDonald and Dr. A Baumgaertner (UC Physics) for providing climate and air pollution data sets. Thanks to the department of Mathematical and Statistics at the University of Canterbury for supporting the travel grant.

6.6. References

Aldrin M, Haff IH (2005) Generalised additive modelling of air pollution, traffic volume and meteorology. Atmos Environ 39: 2145-2155.

Colebrook JM (1978) Continuous plankton records: zooplankton and environments, North-East Atlantic and North Sea. Oceanol Acta 1:9-23.

Fukuda K (2004) New improved methods for application and interpretation of SSA: A case study of climate & air pollution in Christchurch, New Zealand, MSc thesis, University of Canterbury, Christchurch, New Zealand.

Fukuda K (2007) Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution. In Proc. of the 2007 IEEE CIDM 2007, 697-704.

Fukuda K, Hudson IL (2005a) Global and local climatic factors on sulfur dioxide levels: comparison of residential and industrial sites. In Proc. of 20th IWSM, 187-194.

Fukuda K, Hudson IL (2005b) Investigations of short-term (hourly) weather influences on CO, NO, NO_2, PM_{10} and SO_2 Levels in Christchurch, New Zealand, In Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural System, 45-52.

Fukuda K, Takaoka T (2007) Analysis of Air Pollution (PM_{10}) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm, In Proc. of the 2007 ACM SAC 2007, 153-157.

Fukuda K, Pearson PA (2006a) Investigation of Singular Spectrum Analysis and Machine Learning for Road Sign Location. In Extended Abstracts of 7th DAS 2006, 29-32.

Fukuda K, Pearson PA (2006b) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-6.

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, Boca Raton.Scott A, Gunatilake M (2004) 2002 Christchurch inventory of emissions to air (R04/03), Environment Canterbury, Christchurch, 2004.

Koening JQ (2000) Health effects of ambient air pollution, how safe is the air we breathe?. Kluwer Academic, Boston.Spate JM, Gibert K, Sànchez-Marrè M, Frank E, Comas J, Athanasiadis I, Letcher R (2006) Data mining as a tool for environmental scientists. In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-22, Burlington.

Kossmann M, Sturman AP (2004) The surface wind field during winter smog nights in Christchurch and coastal Canterbury, New Zealand. Intl J Climatol 24:93-108.

Kumaresan R, Tufts DW (1980) Data-adaptive principal component signal processing. In Proc. of 19th IEEE Conf. on Decision and Control, 949-954.Li S-T, Shue L-Y (2004) Data mining to aid policy making in air pollution management. Expert Syst Appl 27: 331-340.

Li T, Li Q, Zhu S, Ogihara M (2002) A survey on wavelet applications in data mining. ACM SIGKDD Exploration 4: 49-68.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2nd edn.

Study II. Data mining and image segmentation approaches for classifying defoliation in aerial forest imagery (Fukuda and Pearson 2006a,b)

6.7. Introduction

This study covers the brief concept of SSA to process the input data for an aerial image processing problem, classifying different levels of defoliation and different landscape types using the C4.5 algorithm. Results from this approach were then compared with a simple clustering algorithm that was also developed in this study.

The increasing availability of remote sensing and geographic data helps in monitoring and management for maintaining the health of forest ecosystems, which is important for the protection of natural resources and the economy. Satellite imagery, a remote sensing technique, is convenient for large-scale surveys, and has been used widely for land cover and habitat mapping using different applications (Friedl and Brodley 1997; Kobler et al. 2006), but it has low resolution and it can be expensive to obtain timely imagery. Alternatively, aerial photography can provide higher resolution to allow monitoring of forest health and identification of tree species at an acceptable level of accuracy (Haara and Nevalainen 2002). White et al. (2005) investigated an automated interpretation method for detecting the *red attack* stage of trees attacked by the mountain pine beetle using satellite imagery, using aerial imagery for validation. However, not all studies are able to access such high quality data. In fact, environmental studies often deal with incomplete or poor quality data, as it is costly to obtain high quality data, and measurement relies on human observations that may be imprecise or uncertain. Hence, methods for processing poor quality data, perhaps involving statistics or knowledge discovery, can be advantageous.

The central interior region of British Columbia has suffered from increasing populations of mountain pine beetle (*Dendroctonus ponderosae*) since 1994 (White et al. 2005). The British Columbia Ministry of Forests and Canadian Forest Service (BCMF and CFS) carries out annual defoliation surveys, where observers in small aircraft sketch infested regions on forest maps. Aerial surveying is said to be “not an exact science…as no matter what type of aircraft, the flying height, the weather, the survey map base, or the biological window, the survey is always going to be less than perfect” (BCMF and CFS 2000). The survey accuracy depends on the observers’ knowledge of the local forest and pests. Usually only estimates of current tree mortality are indicated, but experienced personnel can estimate damage intensities fairly accurately with help of a multi-stage sampling procedure including aerial

photography, GPS point and ground plot data, to ensure accuracy by enabling cross-validation.

Decision tree algorithms are considered suitable for remote sensing applications, since they are flexible and robust with respect to non-linear and noisy relations among input features and class labels, and prior assumptions regarding the distribution of input data are not required (Friedl and Brodley 1997). However, the purpose of this study is to develop statistically and computationally driven methods via data mining and image segmentation, to add insight towards *aerial imagery interpretation* for the annual defoliation survey procedure. Typically, classification accuracy is tested in data mining projects by cross validation on as much data as possible, but this study takes a different approach. Data mining is used for knowledge discovery: the extent of infested regions and land cover are predicted using a decision tree classifier, C4.5 (Quinlan 1993), based on the contents of *only a few known* (training) data points that have been manually pre-identified by an expert. Then, the classification tree is created from a small proportion of the data (only a few known data points) and tested on the rest of the data to model the intended use of the system: for estimating tree mortality and land cover when complete ground truth is not available. The available image for this study has only low resolution (287 × 313 pixels), uneven lighting and varying scale, so the data mining approach is designed to be applicable to low quality imagery. It identifies patterns directly using the training data, thus traditional image pre-processing to normalize the image or remove noise is unnecessary.

In comparison to the data mining approach in this study, a simple single linkage non-hierarchical clustering image segmentation method is designed for this study. This method uses manually-created pixel classification functions to detect attacked trees, then clusters pixels into regions, and estimates the tree mortality density in each region. A purpose is to demonstrate how the pixel classification can be different from the tiled approach classification.

6.8. Defoliation imagery

Aerial imagery (Fig. 6-4) was captured in Flathead Valley, Nelson Forest Region, in British Columbia, Canada, which suffers from mountain pine beetle attack. The studied aerial imagery is a low-resolution photo (287 × 313 pixels), downloaded from the source website (BCMF and CFS 2006) and the use of this imagery was authorized by B.C. Ministry of Forests and Canadian Forest Service.

Over the mountain pine beetles' one-year life cycle, tree foliage becomes chlorotic, then yellow, and finally fades to red. The BCMF and CFS (2000) define three levels for tree

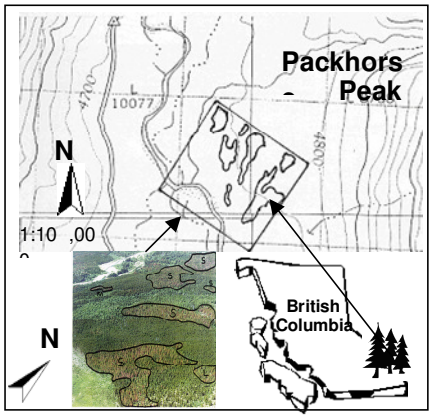


Fig. 6-4 Location of the aerial image site, Flathead Valley, Nelson Forest Region, and original 70 mm photo (BCMF and CFS 2000)

Table 6-3 Tree mortality and land cover classification criteria.

Tree mortality Classification	Criterion	
	Bark beetles	Defoliators
Severe (S)	>30% of trees recently killed	Bare branch tips and completely defoliated tops. Most trees sustaining more than 50% total defoliation.
Moderate (M)	11-29% of trees recently killed	Pronounced discoloration. Noticeably thin foliage. Top third of many trees severely defoliated. Some completely stripped.
Light (L)	1-10% of trees recently killed	Discoloured foliage barely visible from the air. Some branch tip and upper crown defoliation.
Land cover classification	Criterion	
Vegetation (V)	Green regions that do not contain trees.	
Ground Surface (Surface)	Regions where the ground surface is exposed.	
Non attack (Non)	Regions that are not included in tree mortality classifications, assumed to be non-attack regions.	

mortality caused by defoliators and bark beetles: severe (S), moderate (M), and light (L). This study adds extra classes for land cover: vegetation (V), ground surface (Surface) and non attack (Non); all classes are shown in Table 6-3. Fig. 6-4 shows one L, one M and five S regions, identified by BCMF and CFS.

6.9. Methods

6.9.1. Data mining approach

To convert the image into a form suitable for analysis, it is divided into relatively large (20 × 20 pixel) tiles (details in the next section). This tile size reduced noise in histograms and represented relevant region characteristics better than smaller (10 x 10 pixel) tiles. Next, training data points are created from the peak values of smoothed histograms of red (R), green (G) and blue (B) colour channels, and their average (A). The histograms are smoothed by Singular Spectrum Analysis (SSA) (Golyandina et al. 2001; Fukuda and Pearson 2006a,b), which was found to provide better results than a Fourier transform low-pass filter. The analysis is improved by adding the difference between each pair of colour peak values, e.g., R-G, to each training data point. Lastly, a decision tree is generated via WEKA and tested using three different sets of training data points to predict the rest of the imagery, followed by stratified cross-validation on the entire image (details in the following section). The predicted classes are then overlaid on the image, to provide visual feedback on the classification results

6.9.1.1. Extraction of histograms

Let $L = \{(n, m), n = 1, \ldots, N, m = 1, \ldots, M\}$ be a 2D lattice of pixels for an image, I, where n and m represent columns and rows respectively. The image, I, is divided into $S = \{(n/p, m/p), 1 \le p \le n, 1 \le p \le m\}$ tiles of $p \times p$ pixels. Here, I is defined by N=313, M=287 with

p=20 (20 × 20-pixel tiles) to give S = (15, 14), a total of 210 regions. Colour frequency histograms H_R, H_B, H_G and H_A are extracted from the four colour channels in each S_p tile. Now, SSA (Golyandina et al. 2001) is applied to smooth each histogram. Each H is treated as a 1D series of length Q=256, $H = (f_0, \ldots, f_{Q-1})$, and transferred into a set of *W-dimensional lagged vectors*, $X_i = (f_{i-1}, \ldots, f_{i+W-2})^T$, where $1 \leq i \leq K = Q - W + 1$ and W is the *window length* ($W \leq Q/2$); for this analysis, W=32. This procedure turns the H series into the *W-trajectory matrix*, $X = [X_1 : \ldots : X_K]$), which can be rewritten as

$$X = (x_{ij})_{i,j+1}^{W,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{W-1} & f_W & f_{W+1} & \cdots & f_{Q-1} \end{pmatrix} \quad (i+j = \text{const}). \tag{6-11}$$

The obtained trajectory matrix, in equation 6-1, is decomposed by the singular value decomposition (SVD) to provide i eigentriples (consisting of eigenvalues, principal components and right singular vectors). The eigentriples are grouped and reconstructed to form the smoothed histograms for each tile in S. Here, the first three eigentriples were used, to provide >75% of the original variance.

Let h represent the smoothed histograms and h_c = (in turn) h_R, h_G, h_B, h_A for each colour channel. Maximum values of each h_c are then calculated for constructing training data points. The differences between each pair of values. e.g., $maxh_R - maxh_G$, are added to increase the number of attributes available for data mining.

6.9.1.2. Singular spectrum analysis for classification process

The J4.8 classifier from WEKA 3.4 (Witten and Frank 2005), based on the C4.5 algorithm (Quinlan 1993), is used to generate decision trees from a small number of training data points to predict tree mortality and land cover (class) for the rest of the image. Note that regions labelled as S are divided into five regions: S1 to S5, in decreasing order of size. Four experiments were performed, selecting training data with three different methods.

1) Manually selected training data: To test if the patterns of colour channel peaks effectively represent tree mortality, tiles were examined to find similar patterns of $maxh_c$ values (*colour patterns*) and verify the connection between these colour patterns and tree mortality/land cover classes, then up to four of the most representative tiles in each class were manually selected as training data points, to produce a decision tree which was tested on the remaining data.

2) Randomly selected training data: To model the real-world training data selection process, first two, then three training data points were selected randomly from each class (S1 to S4, M, L, Surface, V and Non) to produce a decision tree, which was tested on the remaining data.

3) Stratified cross-validation: To test the overall performance of the decision tree method, the entire dataset was tested using ten-fold stratified cross-validation, with S1-4 combined into a single S class.

Note that the S5 region is ignored as it only contains one tile, and L and M are also small, with four and two tiles respectively. The predicted class for each tile is overlaid on the imagery to allow visual interpretation of classification results (except for cross validation). To reduce the visual complexity of result images, the classes used internally are reduced to S, M, L, V, Surface and Non. Classification accuracy is presented as four numbers. *Overall accuracy* is the proportion of correct classifications when decision trees are tested on the entire dataset, including the training data used to create them. *Excluding training set* is the proportion of correct classifications when the training set is excluded from the test set. *Weighted* values weight different errors differently, giving a greater penalty for larger errors, e.g., Non misclassified as *S*, than for errors between adjacent classes, e.g., L misclassified as M.

6.9.2. Image segmentation approach

The image segmentation approach, in contrast to the single-pass tile-based data mining method, first attempts to detect whether individual pixels belong to attacked trees, then groups these *attack pixels* into regions, and finally quantifies the severity of the attack in each region.

Let α represent the source image, such that $\alpha(x, y)$ represents the pixel at column x and row y of the image. Let $\alpha_H(x, y)$, $\alpha_S(x, y)$ and $\alpha_V(x, y)$ represent the hue, saturation and value attributes of the pixel $\alpha(x, y)$. Hues lie in the range $[0^o, 360^o)$, while saturations and values lie in the range [0%, 100%].

6.9.2.1. Pixel classification

First, pixels are classified as to whether they are expected to correspond to attacked trees, using one of a number of manually-designed classifiers. The seven classifiers, with different hue, saturation and value criteria, are shown in Table 6-4.

Equation 6-12 shows how this step produces a *detection matrix*, *D*, for the *A* classifier:

$$D(x,y)=\begin{cases} \alpha_H(x,y)<24^\circ \ \& \\ 1 \quad \alpha_S(x,y)>20\% \ \& \\ \alpha_V(x,y)>50\% \\ 0 \quad \text{otherwise} \end{cases} \tag{6-12}$$

6.9.2.2. Region detection and tree mortality quantification

Next, a new matrix, E, is created. Each cell in E contains the sum of all values within 10 cells of the corresponding value in D: $E(x,y) = \Sigma D(x',y')$, $\forall \sqrt{(x-x')^2+(y-y')^2} < 10$. A threshold, τ, is defined as 10% of the maximum value in E: $\tau = \max(E) / 10$. This threshold is then applied to E to produce R, which is equal to 1 where the corresponding element in E is greater than τ.

$$R(x,y)=\begin{cases} 1 & E(x,y)\geq\tau \\ 0 & E(x,y)<\tau \end{cases} \tag{6-13}$$

Now, a connected component analysis is performed to extract connected groups of nonzero elements in R, which represent possible infested regions. Let $C_T(n)$ represent the count of pixels in the n^{th} connected component, and $C_A(n)$ the count of detected *attack pixels* in the component (i.e., pixels for which $D(x, y) = 1$). Let $C_R(n)$ equal the proportion of attacked pixels in region n.

$$C_R(n)=\frac{C_A(n)}{C_T(n)}. \tag{6-14}$$

Regions are classified as severely attacked if $C_R(n) \geq 0.3$, moderately attacked if $C_R(n) \geq 0.2$, lightly attacked if $C_R(n) \geq 0.1$, and non-attack otherwise.

6.10. Results and discussion

6.10.1. Patterns of colour channel histogram peaks

Plotting colour channel histogram peaks for manually selected similar tiles (Fig. 6-5) identified nine distinct colour patterns from the six classes (S, M, L, Non, V and Surface).

Table 6-4. Pixel classification methods.

Method	Hue criterion	Saturation criterion	Value criterion
A	H < 24°	S > 20%	V > 50%
B	H < 54°	S > 10%	-
C	H < 54°	S > 20%	V > 50%
D	H < 24°	S > 20%	V > 50%
E	H < 24°	S > 20%	V > 39%
F	-	S < 10%	-
G	245° < H < 305°	S > 20%	V > 50%

The Non tiles contain three distinct patterns (Fig. 6-5 a, b, and g), visible in the image as yellow green, light green and grey. S tiles show two patterns: reddish yellow (S1 and S2, Fig. 6-5 e) and grey (S3 and S4, Fig. 6-5 f).

These results reflect the heterogeneous nature of Non and S regions. Interestingly, the use of only four values from each relatively large tile successfully describes the colour changes that take place over time during the process of tree *infestation* and *mortality*. The sequence may start from *light green* Non (Fig. 6-5 a), which has a high green peak, and red and blue peaks at zero. Next, red is added, as seen from *yellow green* Non (Fig. 6-5 b), which overlaps with L (Fig. 6-5 c). Then, blue is added as the class changes to M (Fig. 6-5 d). During the S stage, the blue peak drops to zero, and the red peak becomes higher than the green peak (Fig.

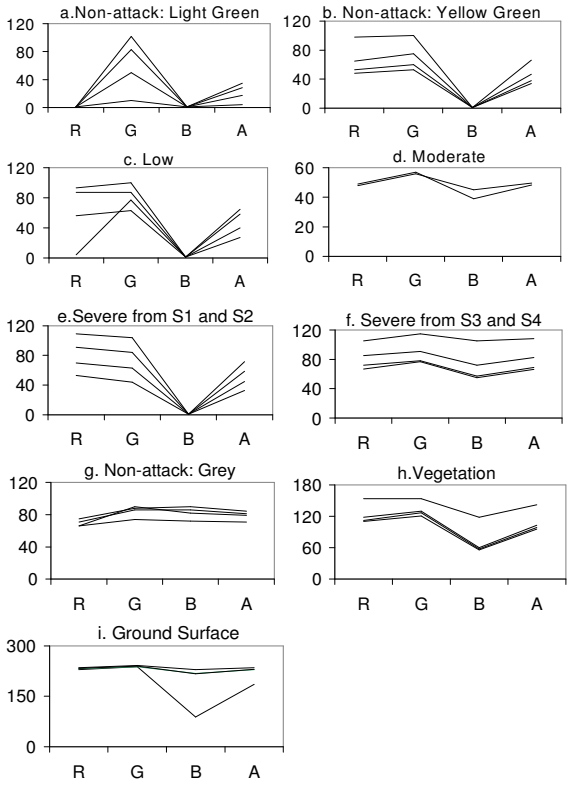


Fig. 6-5 Colour channel histogram peaks, showing distinct patterns.

Table 6-5 Confusion matrices for test results.

Cross validation on full image							
	Actual value						
Detected	S	M	L	V	Surf	Non	
S	**23**					7	
M							
L							
V				**11**			
Surface					**4**		
Non	33	2	4	2		**154**	
Training accuracy							80.0%
Weighted accuracy							82.2%
With cross-validation							**74.6%**
Weighted x-validation							**79.4%**

Colour pattern analysis							
	Actual value						
Detected	S	M	L	V	Surf	Non	
S	**33**					20	
M		**2**				6	
L	15		**4**			49	
V	1			**13**		9	
Surface					**4**		
Non	7					**77**	
Overall accuracy							55.4%
Weighted accuracy							75.7%
Excluding training set							**49.0%**
Weighted w/o train set							**72.0%**

Three training data points							
	Actual value						
Detected	S	M	L	V	Surf	Non	
S	**35**		3			45	
M	1	**2**				1	
L	5		**3**			26	
V				**6**		5	
Surface			4		**4**	4	
Non	15		1			**80**	
Overall accuracy							54.2%
Weighted accuracy							68.2%
Excluding training set							**49.1%**
Weighted w/o train set							**64.8%**

Two training data points							
	Actual value						
Detected	S	M	L	V	Surf	Non	
S	**38**					82	
M		**2**				11	
L	6		**2**			27	
V				**9**		5	
Surface				4	**4**	4	
Non	12		2			**32**	
Overall accuracy							36.3%
Weighted accuracy							52.1%
Excluding training set							**31.1%**
Weighted w/o train set							**48.2%**

6-5 e). Red, green, blue, and average peaks are similar in S3/S4 (Fig. 6-5f), as well as *grey* Non (Fig. 6-5 g), appearing flat when plotted, although S3/S4 tiles have a slightly lower blue peak.

These tiles only appear in the top quarter of the image, suggesting that the greyness may be due to light conditions and distance from the photographer, although another possibility is a high concentration of long-dead trees, which are known to appear grey, but marked Non because only recent attack is labelled (BCMF and CFS, 2000). Structures of M (Fig. 6-5 d), V (Fig. 6-5 h), S3/S4 (Fig. 6-5 f) and Surface (Fig. 6-5 i) have similar patterns, but different ranges of peak intensities. There is an overlap between Surface and V, which may cause some misclassification. As S and Non tiles are heterogeneous due to poor lighting in the imagery, training data sets must contain samples from each different variant of S and Non for classification to be successful.

6.10.2. Classified imagery and confusion matrices

Classification figures of accuracies and confusion matrices are shown in Table 6-5, and images with overlaid prediction results are shown in Fig. 6-6.

The classification accuracy on the test set is 75% for cross validation and 31-49% when using smaller training data sets. Weighted accuracy figures are 79% for cross validation, 72% for manually selected training data and 52-68% for randomly selected training data. The best data mining results were obtained using the full image data, with all classes classified well except S (41% recall). The next best results were from the colour channel pattern analysis (Fig. 6-6, A), which detected V and Surface correctly (100%), although S (59%) was often misclassified as L or Non, and Non (48%) was often misclassified as L or S. This could be due to the similarity between S1/S2, L, and *yellow green* Non classes (the classifier selects the *middle ground* class L between the extremes of S and Non), or perhaps the gaps between large S1/S2 regions are bridged by L, as also seen from one space between an M and an S3/S4 region classified as M, but further investigation is required. As training tiles are added from the top of the image (especially *grey* Non, Fig. 6-5 g, and S3/S4, Fig. 6-5 f), misclassification of Non and S in that region is reduced. For the same reason, V and Surface tiles had 100% recall in this test.

The test using two training data points (Fig. 6-6, C) classified Surface (100%) and V (69%) tiles satisfactorily, though S (68%) tiles were often misclassified as Non and most Non tiles were misclassified. Classification improved when three training data points were used (Fig. 6-6, B), with much better separation of S and Non regions (Non recall improved from 20% to 50%).

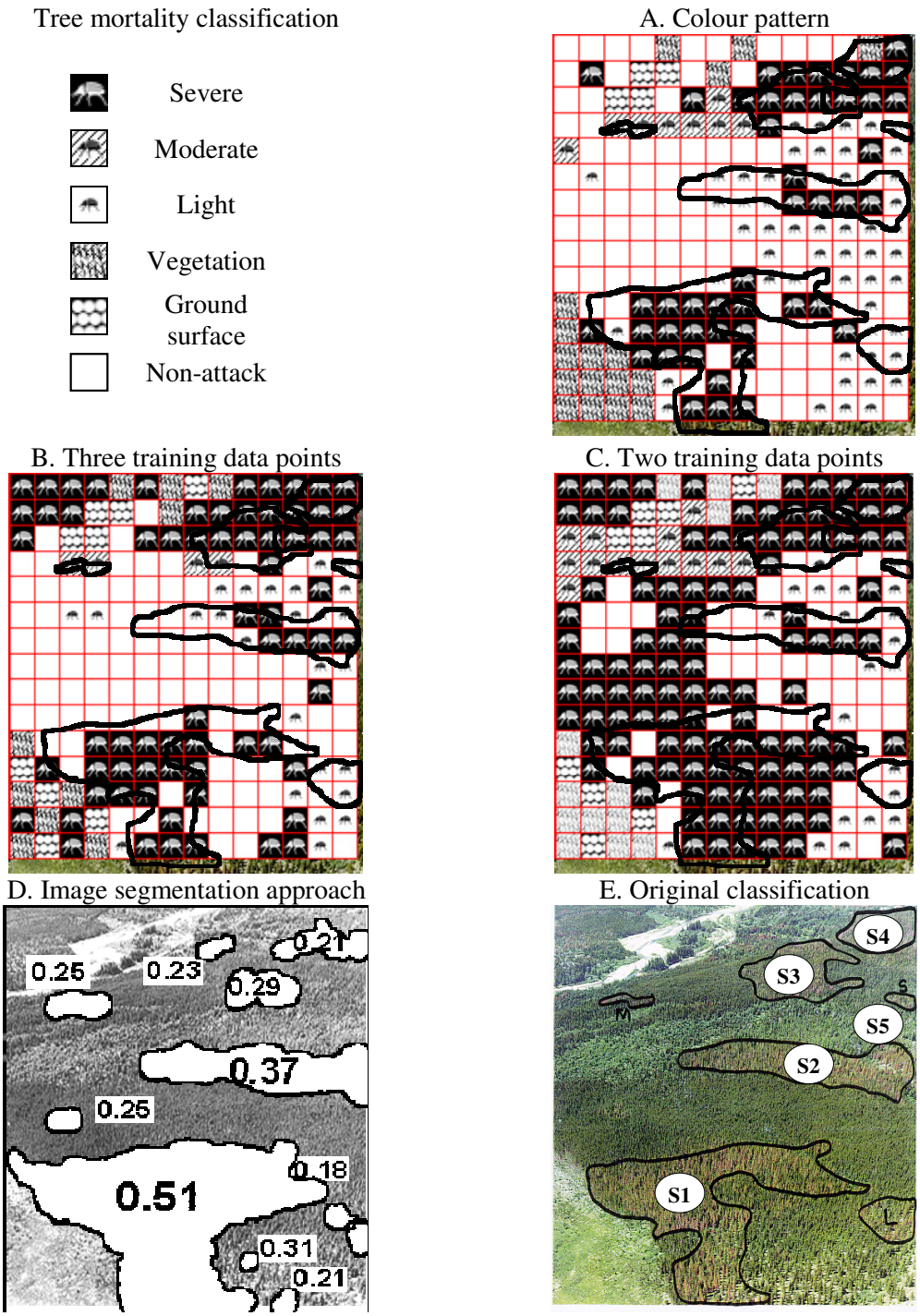


Fig. 6-6 Prediction result overlay images.

Manually classified regions of tree mortality from E appear as thick outlines in A-C.

Both results classified both top corners, which appear grey in the original image, as S, and the three-point analysis confused several S3/S4, V and Surface tiles. This may be due to the similarities in their colour patterns, as previously discussed. Performance on L and M tiles was poor, with only training data classified correctly.

Overall, all tests had significant misclassification between S and Non due to the heterogeneity of Non regions and aerial photography conditions, as previously discussed.

Performance on small training sets still needs to be improved, although encouraging results were achievable with well identified heterogeneous training data. The image segmentation approach (Fig. 6-6) detected regions marked by the BCMF and CFS (2000) as infested (S, M, L classes) with 84% accuracy, and identified three regions that had not been flagged by the human observer but appear infested on the image. The only serious misclassification occurred in the top right-hand corner of the image, where two large S regions were classified as M and only partly located.

Additionally, Fig. 6-7 shows an example of decision tree structure, taken from the best decision tree for colour pattern analysis, with 9 leaves and a size of 17. For classifying the aerial imagery, the blue histogram peak appears most important, followed by red and green. Grey (A) peaks were not found to be important alone, but R-A (red peak minus grey peak) was used to distinguish between Non, M and S.

6.11. Conclusions

Results with small training data sets need to be improved, although the rate at which classification accuracy improves with the addition of well identified heterogeneous training data is encouraging for further investigation. In future, more image features will be considered, and techniques such as hybrid decision trees, which Friedl and Brodley (1997) found to provide higher classification accuracy, will be investigated. Higher quality input data, e.g., higher resolution, overhead angle, even lighting, will improve results. The (generic) data mining approach can be applied to other image classification problems, and will be used to produce better initial pixel classification rules for the image segmentation method, while image analysis techniques will be used to provide richer data points for the data mining approach.

6.12. Acknowledgements

The authors thank three anonymous reviewers for their detailed comments. K. Fukuda acknowledges the University of Canterbury for providing a travel grant for iEMSs 2006 attendance. Thanks to the British Columbia Ministry of Forests and Canadian Forest Service (Dr. M. Wulder) for allowing us to investigate the imagery.

6.13. References

BCMF and CFS (2000) “Overview Aerial Survey Standards for British Columbia and the Yukon” Forest Health Aerial Overview Survey Standards for British Columbia. The B.C. Ministry of Forests adaptation of the Canadian Forest Service’s FHN Report 97-1, 23-24.

BCMF and CFS (2006) Integrated land management Bureau, Available via http://ilmbwww.gov.bc.ca/risc/pubs/teveg/foresthealth/assets/aerial-1.jpg. Accessed on 11 February, 2006.

Friedl MA, Brodley CE (1997) Decision Tree Classification of Land Cover from Remotely Sensed Data, Rem Sens Environ 61: 399-409.

Fukuda K, Pearson PA (2006a) Investigation of Singular Spectrum Analysis and Machine Learning for Road Sign Location. In Extended Abstracts of 7th DAS 2006, 29-32.

Fukuda K, Pearson PA (2006b) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-6.

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, Boca Raton.

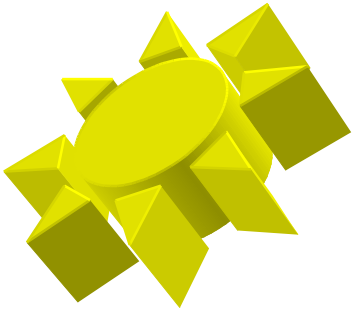
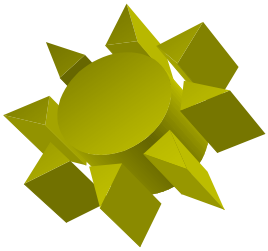
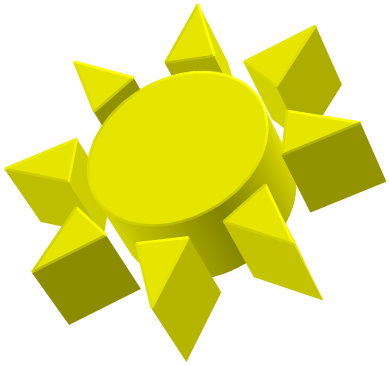
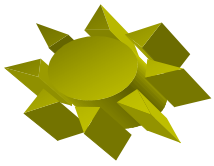
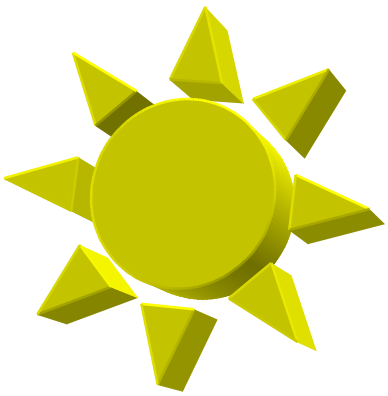
Haara A, Nevalanine S (2002) Detection of dead or defoliated spruces using digital aerial data. For Ecol Manage 160: 97-107.

Kobler A, Džeroski S, Keramitsoglou I (2006) Habitat mapping using machine learning-extended kernel-based reclassification of an Ikonos satellite image. Ecol Model 191: 83-95.

Quinlan JR (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo.

Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2nd edn.

White JC, Wulder MA, Brooks D, Reich R, Wheate RD (2005) Detection of red attack stage mountain pine beetle infestation with high spatial resolution satellite imagery. Rem Sens Environ 96: 340-351.

Chapter 7. Future plans and software development for environmental science problems

This thesis has three fundamental aims and motivations. Firstly, to develop flexibly applicable new tools, TNS and TNS-A, to assist scientists in investigating environmental data. The use of appropriate methods to investigate environmental data will help decision making in the policy development and management process. Secondly, to introduce the use of new, theoretical or unknown computer algorithms, such as the *K*-MSA, by adjusting and maximizing their applicability and practicality to assess environmental science problems to bring new insights. One part of this second aim was to demonstrate the unique advanced statistical and mathematical data pre- decomposition method, SSA, to help obtain improved results with the C4.5 algorithm by pre-processing noisy measurements. Thirdly, to promote, encourage and motivate various environmental scientists to use ideas and methods developed in this thesis. Since each chapter contains its own conclusion section, this chapter discusses overview of methodological conclusions, and introduce how research performed and introduced in this thesis will be used and promoted in the near future, by collaborating with various scientists from various countries, to improve environmental science research.

7.1. Overview of methodological conclusions

This section describes additional future methodological challenges in the conclusion of each chapter. The later sections introduce the outcome of this thesis, including future collaboration and research development.

Chapter 1 described the concept of Knowledge Discovery in Databases (KDD). Throughout this thesis, by applying various available computer algorithms as data mining techniques, e.g., C4.5, the *K*-MSA, Ant-Miner, TNS and TNS-A, the concept of KDD has been shown to be useful and powerful to discover hidden knowledge about data, especially smaller data, which are often not suitable to produce a prediction model, as the sample size limits effective rule extraction. In data mining applications, it is unconventional to attempt to statistically quantify the knowledge discovery investigation results, e.g. by p-value, since knowledge discovery investigation generally extracts rules from a single set of observational data, with 10-fold cross validation applied to obtain a classification accuracy to assess the quality of the rules, whereas prediction models test the extracted rules on a validation set. It may be a future challenge to develop or apply a comparable method in statistics to assess the quality of the extracted rules even for knowledge discovery investigations.

Chapter 2 introduced a new attribute selection method, Tree Node Selection (TNS), and its benchmarking experiment (Fukuda and Martin in press). TNS was found to produce the most consistent results in attribute reduction and classification accuracy improvement out of five well known attribute selection methods. Chapter 3 described a new assessment tool for decision tree structure, Tree Node Selection for assessing decision tree structure, TNS-A. Both TNS and TNS-A were applied as knowledge discovery tools to understand the structure of the sea container contamination pathway and the Weed Risk Assessment model (WRA) by extracting important attributes, i.e., factors or questions(Fukuda and Brown 2007a,b). It will be interesting to use TNS as an assessment tool to compare different decision tree learning schemes, e.g., lookahead (details in Chapter 2), to examine ranking nodes. We are also working towards publishing a journal paper and plan to release the TNS code in a suitable form to run as part of WEKA. Similarly, TNS-A can be used to assess how different decision tree learning schemes select different nodes and classes.

Chapter 4 introduced a new use of the *K*-MSA and a new parameter, the threshold value (Fukuda and Takaoka 2007a,b). The *K*-MSA was demonstrated on air pollution, climate and health data (Fukuda and Takaoka 2007a). In Chapter 5, the *K*-MSA was compared with the *k*-means clustering algorithm as an alternative clustering method, using a unique benchmark data set (Bumpus sparrow data) and was also applied to investigate the spatial hawthorn

distribution (Fukuda et al., 2008). If the array were large, then clustering or GIS techniques would be more suitable to detect curved regions. It will be desirable in the future to develop a *K*-MSA technique which can detect flexibly defined maximum regions, e.g., circular or curved.

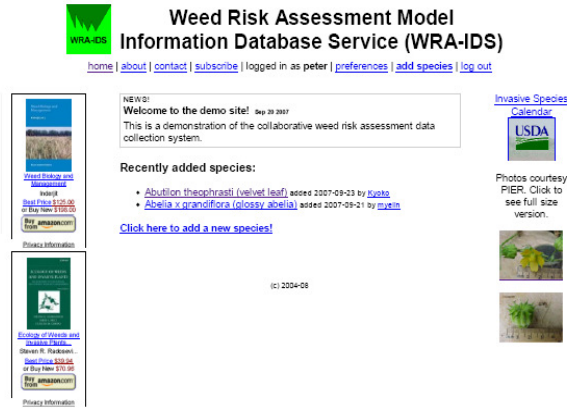
Chapter 6 introduced a new data pre-processing method, Singular Spectrum Analysis (SSA), for a decision tree algorithm, and was applied on air pollution and climate time series (Fukuda 2007), and defoliation imagery (Fukuda and Pearson 2006). It will be interesting to test in future on different learning algorithms, e.g., naive Bayes, to observe how data pre-processing can be effectively applied.

7.2. Future plans for TNS and TNS-A for the Weed Risk Assessment model

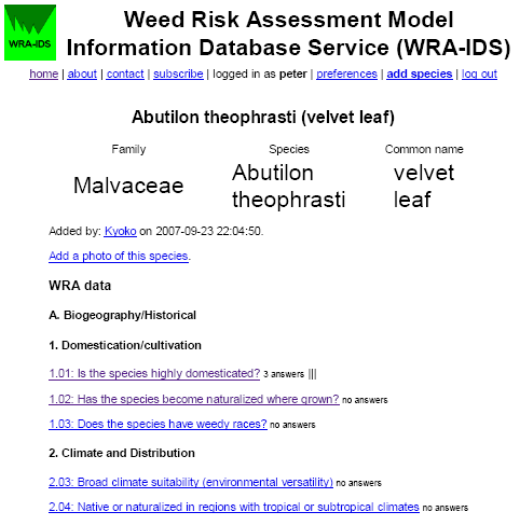
Even though the Weed Risk Assessment (WRA) model is not a part of the legal process, many countries, such as Australia, New Zealand, the USA, Japan and South Africa, have found that the model helps the decision making process for assessing the impact of invasive plants. Dr. Nishida at the Japanese National Institute for Agro-Environmental Sciences (NIAES) and I discussed in January 2008 how we could improve the Weed Risk Assessment (WRA) model procedure for Japanese weeds, and improve communication among scientists to help the WRA procedure. We met during the International Conference on Ecology and Management of Alien Plant Invasions in 2007, EMAPi9 (Fukuda and Brown 2007a). Dr. Nishida has been promoting the effective use of the WRA model in Japan to supplement the Japanese Invasive Alien Species Act, which commenced in 2005 (see details in Nishida et al. 2008). At the conference, Dr. Nishida and her colleagues, and other scientists from various countries, raised some issues about accessibility of alien plant information among scientists.

In order to assess whether a given plant will be a future threat, each country needs to access as many resources as possible during the decision making process. However, the available information for the alien species is often written in its native region's language, e.g., a plant native to Japan may be well documented in Japanese, which may be difficult to access or read for an assessor in Chile. Alternatively, it can be difficult to locate experts on particular plants. Assessors can take many hours to months to find the small piece of relevant information about a particular plant to answer a WRA question.

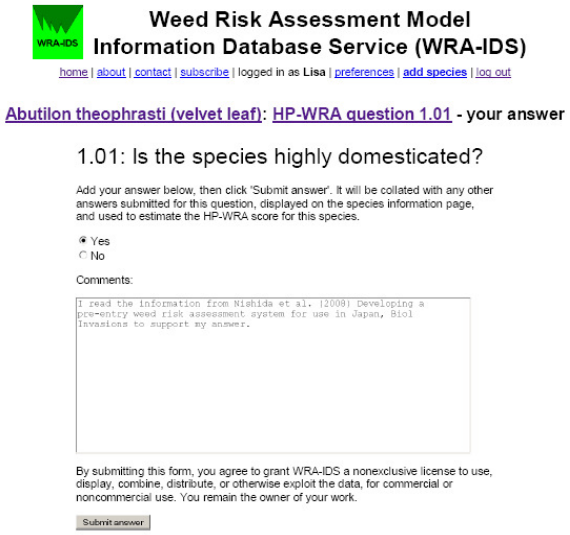
A. The WRA-IDS home page



B. The WRA questionnaire page.



C. Comment function and the WRA questionnaire decision page.



D. Sharing thoughts and comments.

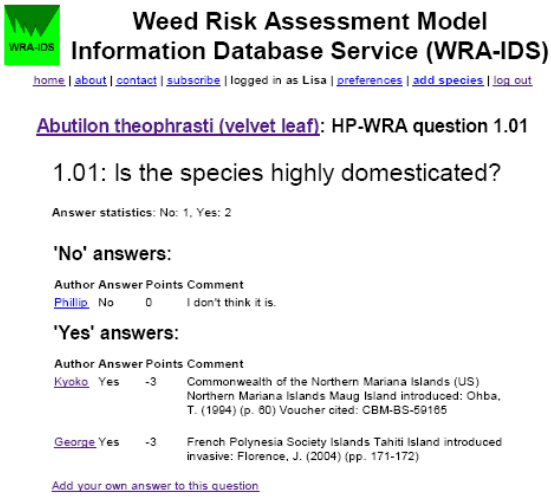


Fig. 7-1 The prototype Weed Risk Assessment Model Information Database Service (WRA-IDS) website.

From further discussions at NIAES[2], I have developed a prototype website, the Weed Risk Assessment Model Information Database Service (WRA-IDS). This website can link and gather information to help with the WRA model, and is shown in Fig. 7-1. This is similar to the Pacific Island Ecosystems at Risk project (PIER), an online archive tool to help the WRA model process by allowing us to share the information about listed plants.

[2] Phytoremediation of land contaminated by heavy metals: copper concentrations in self-sown tomato plants from a landfill in New Zealand" (Fukuda and Taylor 2005), and "Investigation of The Weed Risk Assessment Model Using Data Mining" (Fukuda and Brown 2007a,b) were presented at the January biodiversity division seminar, the National Institute for Agro-Environmental Sciences, Biodiversity Division, Ibaragi, Japan, organized by Dr. Nishida (9 Jan, 2008).
http://www.niaes.affrc.go.jp/rplan/library/seminar/info0801.html

PIER is a searchable dictionary of plants, which archives information about many plant species, along with their assessment scores and details from the WRA model (PIER 2008). The proposed WRA-IDS would archive information (Fig. 7-1) have a similar search function to PIER, but the WRA-IDS would be backed by a database with a comment (blog) like function, i.e., Fig. 7-1, A shows the WRA-IDS home page and Fig. 7-1, B demonstrates using an example plant, velvet leaf. This blog function would allow any user to add comments such as thoughts on the WRA decision, and references to suggest evidence about the plant (Fig. 7-1, C).

The usefulness of the proposed WRA-IDS is that if you were to search for a particular species and open the particular WRA question, you could identify who holds useful information on that species. For example, and Fig. 7-1, D, shows that George answered "*no*" to question 1.01 of the WRA for velvet leaf, with a reference. This would help users identify who has quality information about plant species of interest, and would allow contacting such people through a WRA-IDS email system. If users updated plant information by adding comments and evidence, the WRA-IDS would act as a large data source to help the decision making process for plants and would help ease the process for the WRA model.

Dr. Nishida and I may seek future funding to expand the concept and develop the WRA-IDS website further. In the near future, I am planning to add new tools, such as simple statistics on the WRA decisions to inform how many people answered *yes* or *no* to each question. It will be also interesting to incorporate TNS and TNS-A methods online. If the user wishes to identify important WRA questions using archived data, the website will run TNS or other attribute selection methods to identify key questions for the particular plant. The global goal from this work will be to provide a useful tool for volunteers to professional researchers to help with assessing the invasiveness of plant species, and hopefully bring the joy of finding information on unknown plants in our community throughout the world.

7.3. Future plan for TNS and TNS-A for the sea container contamination risk profiles

The pre-detection of risky goods using data mining techniques may be possible with information provided from the documentation of the goods – certificate of goods and Ministry of Agriculture and Forestry (MAF) quarantine declarations for imported containers – prior to the ship entering the country. Taylor et at. (2000) suggested that we need a first line of defence that includes 'early warning' systems, intelligence-gathering and information-sharing involving international co-operation, since we cannot completely rely on border inspection of goods and passengers entering New Zealand. A prototype early warning system,

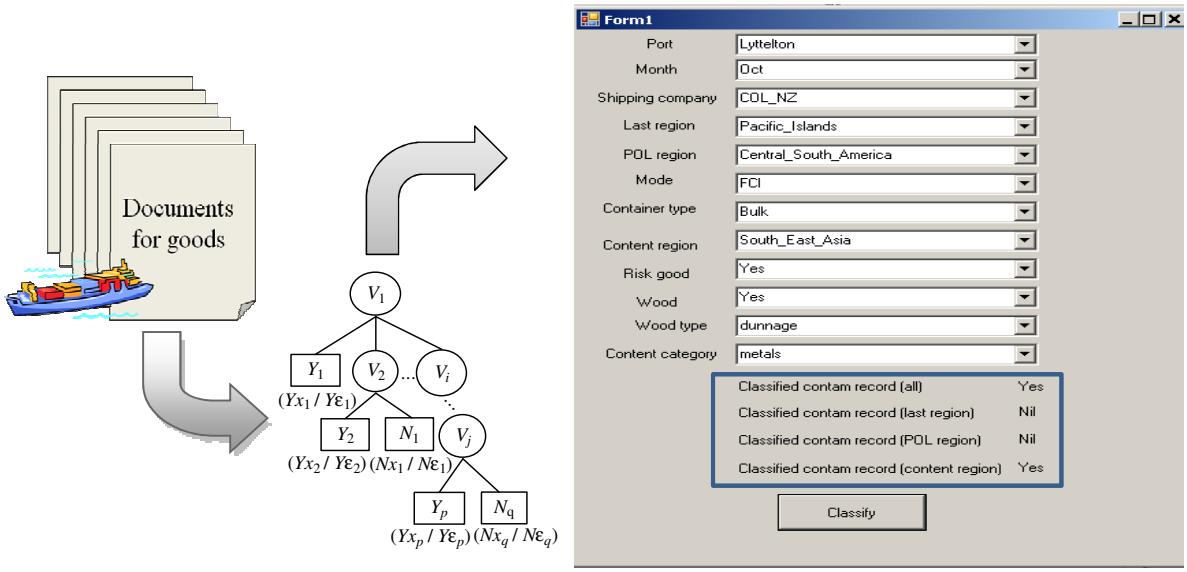


Fig. 7-2 Prototype early warning system suggested to MAF, using documents for goods to generate the decision tree via TNS and TNS-A to predict potential sea container contamination.

A box indicates the various prediction categories, e.g., classified contamination record (prediction) using all input attributes (risk profiles) for the contamination, *yes* or *no*.

built with C4.5, TNS and TNS-A, was developed in this thesis to help predicting the contaminated sea container, in Fig. 7-2. This idea and software were introduced to the Ministry of Agriculture and Forestry (MAF) New Zealand via a seminar, organized by Dr. Whyte[3]. Since then, Dr. Whyte and her team have experimented with data mining techniques, and commented that such a proposed early warning system would be advantageous for MAF inspection process and could be a considerably useful tool.

The use of early warning tools helps to guide where and which containers to examine even before their arrival. For example, the risk profiles for the single container, e.g., port and container type in Chapter 3, Study II, were obtained from the container certificate, and were used as input data to generate the decision tree. Entry of such input information can be carried out via scanning the documentation bar code on the shipping certificate, as the bar code could be issued when the trader entered their information electronically, shown in Fig. 7-3. The generated decision tree can be further improved by updating with new input, as available. The prediction ability could also be improved by incorporating the most predictive input variables (attributes), identified via TNS and TNS-A, which allow changing the weight values for the input attributes. The outputs of the decision tree are the decision and a probability of whether the container would be contaminated or not. For example, if a container is found to have a high probability of contamination, e.g., 0.80, MAF could

[3] "Using data mining techniques for sea container risk analysis" was presented at the Ministry of Agriculture and Forestry for the Biosecurity New Zealand Data Analysis team in Auckland, organized by Dr. Whyte (3 Aug, 2006).

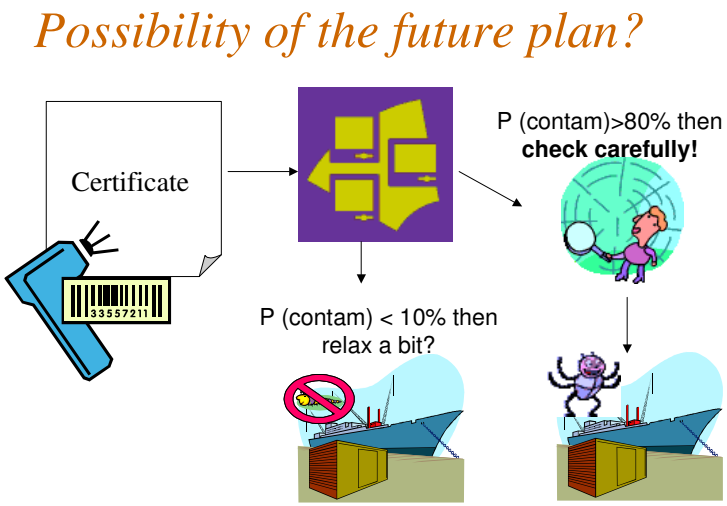


Fig. 7-3 Possible suggestion for the early warning system for the sea container contamination detection using a data mining prediction tool.

investigate that specific container very carefully, as suggested in Fig. 7-3. However, this does not mean that the investigation process fully relies on the detected results from the software. Obtained information from the software will be used to provide a cost effective detection method, in terms of time, by suggesting or pointing out potentially risky containers. Dr. Whyte has provided a larger data set to experiment further with TNS and TNS-A methods, which allows me to keep investigating the sea container contamination problem, to construct improved decision tree structures, and help provide knowledge for the future biosecurity policy strategy.

7.4. Future plan for the *K*-MSA as a tool for GIS software

Another goal of this thesis was to introduce the newly developed *K*-MSA (eg., Bae and Takaoka 2006) as a knowledge discovery tool for various environmental science problems and scientists[4]. During the International Congress on Modelling and Simulation, MODSIM 07 (Fukuda and Brown 2007b; Fukuda and Takaoka 2007), Dr. Wieland from the Leibniz Centre for Agricultural Landscape Research (ZALF) in Germany and I discussed a future collaboration to produce a more practical and applicable solution for the *K*-MSA method for environmental scientists by implementing it into GIS software. Dr. Wieland and his colleagues have developed the integrated Spatial Analysis and Modeling Tool, SAMT (see

[4] Prof. Takaoka presented "Investigation of the maximum association for suicide rate and social factors using computer algorithm (Fukuda and Takaoka 2007b)" at the 8th RCSS international workshop (28 Nov, 2007) and I presented "Computer algorithm for social simulation based on hospital data (Fukuda and Takaoka 2007a)" at the 10th RCSS international workshop (16 Jan, 2008), organized by Prof. Ukai at the Research Center of Socionetwork Strategies, Kansai University, Osaka.
http://www.rcss.kansai-u.ac.jp/workshop_E.html

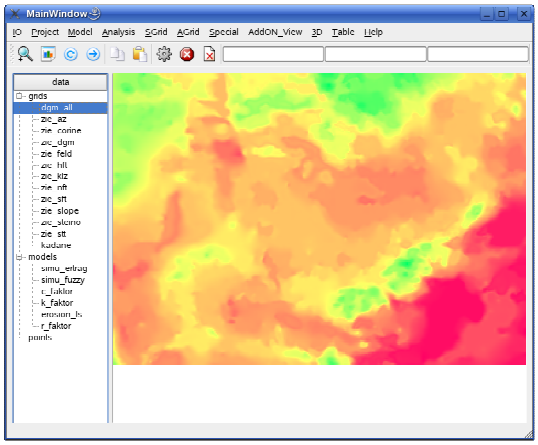


Fig. 7-4 Example of SAMT using the soil and climate index (provided by Dr. Wieland).

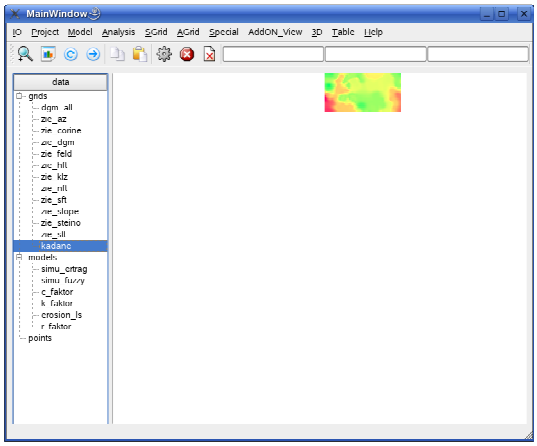


Fig. 7-5 The brightest spot selected by the maximum subarray algorithm for the soil and climate index (provided by Dr. Wieland).

details in Wieland et al. 2006), a free GIS software package that integrates models from different sciences, e.g., economics and ecology, by adjusting and modifying the original methods to be applicable for environmental science applications. Already integrated are fuzzy models and neural networks, and now we are planning to integrate the *K*-MSA. This thesis added two new concepts to the *K*-MSA to make it more applicable to environmental science problems: a new *weight value* parameter and the randomization test (Chapter 5, Study II), but the *K*-MSA is still only available as code in the C programming language. As most environmental scientists may not be familiar with C programming, integrating the *K*-MSA into such software is expected to make it accessible to more users and increase its popularity.

An Example demonstration of the *K*-MSA in SAMT, provided by Dr. Wieland, is shown in Fig. 7-4 and Fig. 7-5, using a soil and climate index. Fig. 7-4 shows the original index and *K*-MSA selected the brightest spot, where was examined as the maximum subarray region (*K*=1), shown in Fig. 7-5.

From here, I will be working as a principal investigator on a project funded by the ISAT Linkages Fund, the Royal Society of New Zealand (awarded in August 2008), with Prof. Takaoka and Assoc. Prof. Brown to collaborate with Dr. Wieland, Dr. Berger and their team at ZALF in Germany in 2008-2009 to promote the *K*-MSA in SAMT using various environmental science applications, e.g., bird and weed spatial distributions.

7.5. Future plan for air pollution, climate and health prediction tool

Lastly, the unique data pre-processing, Singular Spectrum Analysis (SSA), was experimentally applied to help generating the improved C4.5 decision tree (Chapter 6) by removing potential noise from noisy measurements. Studies were conducted from predicting

air pollution levels using various noisy climate measurements and from identifying the different levels of defoliated regions on aerial imagery.

The overall knowledge and methods developed in this thesis will be now incorporated to improve the model of the hospital admission rate to Christchurch Hospital for acute cardio-respiratory conditions using a variety of factors, such as air pollution, climate and virology by means of a hybrid method that incorporates a unique combination of statistical, mathematical and computational algorithms. Developing such a new hybrid model helps identify and understand potential cause and effect relationships between air pollution, climate and human health, whereas air pollution and health studies are generally investigated by statistical time series analysis, e.g., General Additive Models (GAMs), as discussed in Chapter 4. This new project will commence in 2009. I will be working as a principal investigator on a project funded by the Canterbury Medical Research Foundation General Project Grant (awarded in September 2008) with Assoc. Prof. Kingham (University of Canterbury Department of Geography) and Dr. Epton and Dr. Hider (University of Otago, Christchurch School of Medicine & Health Sciences). We believe that the improved model will make differences in our community and our policy making process, as the model can act as an early warning system, like the one suggested to MAF, but in this case to help reduce the acute admission rate by suggesting, e.g., avoiding exposure to outdoor air pollution in advance of predicted high pollution levels. Furthermore, the prediction method developed in the new research is planned to be integrated into the hospital operation system to help plan and organise the automated cost effective hospital operation, e.g., scheduling nurses and numbers of beds to meet needs, in advance. This will help increasing the quality of hospital care that consequently will improve the quality of life in New Zealand, and in future, worldwide.

7.6. Overall conclusions

This thesis focused on developing new tools or introducing both well-known, e.g., the C4.5 algorithm, and new or unknown computer algorithms, e.g., the *K*-MSA, to environmental science problems by demonstrating how they could be used to discover different aspects of information about data. Increasing computer technology allows us to develop many more new computer algorithms and different learning schemes. However, even though there are many techniques available for us to explore, we often do not realize what are available and how to make use of them. Computer algorithms tend to be developed to solve a specific problem arising in theoretical or practical computer science studies. It is not always directly applicable outside the field. In recent years, data mining techniques have become popular among environmental scientists, though statistical approaches are generally well used

to provide quantitative analyses. My challenge and interest is to continue bridging between various environmental science problems and a variety of useful methods developed from different disciplines to help understand our problems using knowledge that I gained from this thesis and studies that were conducted in the past (summary my research contributions are listed in Appendix 7-1). Anything can be helpful and useful, if we know what and how to use it. If not, I would like to keep working hard to make it useful, so many of us can explore to obtain more knowledge about our problems, if that would help us improve our environment that we live in.

7.7. Acknowledgements

Thanks to the thesis examiners, Prof. Ukai and Assoc.Prof. Bardsley, to provide me useful comments to improve my thesis.

7.8. References

Bae SE, Takaoka T (2006) Improved algorithms for the *K*-Maximum Subarray problem. Comput J 49:358-374.

Fukuda K (2007) Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution. In Proc. of the 2007 IEEE CIDM 2007, 697-704.

Fukuda K, Brown J (2007a) Investigation of The Weed Risk Assessment Model Using Data Mining, Intl Conf. of 9th EMAPi9, abstract.

Fukuda K, Brown J (2007b) Classification Rule Extraction by Ant-Miner for Weed Risk Assessment, In Oxley L and Kulasiri D. (eds) MODSIM 07, 2882-2888.

Fukuda K, Brown, J, Williams, P, Kean, J (2008) The *K*-Maximum Subarray algorithm as an alternative clustering analysis for the spatial weed aggregation pattern, NZSA 2008 (abstract).

Fukuda K, Martin B (in press) Decision Trees as Information Source for Attribute Selection, In Proc. of the 2009 IEEE CIDM, 0-8.

Fukuda K, Pearson PA (2006b) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery In: Voinov, A, Jakeman AJ, Rizzoli AE (eds). In Proc. of the 3rd iEMSs, 0-6.

Fukuda K, Takaoka T (2007a) Analysis of Air Pollution (PM_{10}) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm, In Proc. of the 2007 ACM SAC 2007, 153-157.

Fukuda K, Takaoka T (2007b) Investigation of the maximum association for suicide rate and social factors using Computer Algorithm, In Oxley L and Kulasiri D. (eds) MODSIM 07, 1381-1387.

Fukuda, K. Taylor, H.H. (2005) Phytoremediation of Land Contaminated by Heavy metals: Copper Concentrations in Self-sown Tomato Plants from a Landfill in New Zealand. In Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural System, 53-59.

Nishida T, Yamashita N, Asai M, Kurokawa S, Enomoto T, Pheloung PC, Groves RH (2008) Developing a pre-entry weed risk assessment system for use in Japan, Biol Invasions. DOI 10.1007/s10530-008-9340-0.

PIER (2008) Institute of Pacific Islands Forestry Pacific Island Ecosystems at Risk (PIER), Plant threats to Pacific ecosystems, the Pacific Island Committee, Council of Western State Foresters, National Association of State Foresters. Available via http://www.hear.org/pier/index.html. Accessed on 20 Sep. 2008.

Taylor B, Cebbie E, Botherway K, James G, Mormorunni C (2000) New Zealand under siege: A review of the management of biosecurity risks to the environment, Office of the Parliamentary Commissioner for the Environment, Wellington.

Wieland R, Voss M, Holtmann X, Mirschel W, Ajibefun I (2006) Spatial analysis and modeling tool (SAMT): 1. Structure and possibilities. Ecol Inf 1:67-75.

7.9. Appendices

Appendix 7-1 List of my publication.

The following are listed in chronological order.

Fukuda K (2004) New improved methods for application and interpretation of SSA: A case study of climate & air pollution in Christchurch, New Zealand, University of Canterbury, Christchurch, New Zealand (thesis).

Hudson I, Fukuda K, Keatley M (2004) Detecting underlying time series structures and change points within a phenological data set using singular spectrum analysis (SSA), IIXth International Biometrics Conference (IBC 2004), pp 167, 11-16 Jul., Cairns (abstract).

Fukuda K, Hudson IL, Pearson PA (2004) Singular spectrum analysis combined with an enhanced Fourier expansion (EFE) method for detecting underlying time series structures – A case study of the impact of notable global and local weather events on the level of air pollution in Christchurch, NZ, ASA Computational Environmetrics Conference, 21-23 Oct., Chicago (poster).

Fukuda K, Hudson IL (2005) Global and local climatic factors on sulfur dioxide levels: comparison of residential and industrial sites, In Proc. of 20th International Workshop on Statistical Modelling (IWSM), pp. 187-194, 10-15 Jul., Sydney (peer reviewed).

Fukuda K, Hudson IL (2005) Investigations of short-term (hourly) weather influences on CO, NO, NO_2, PM_{10} and SO_2 Levels in Christchurch, New Zealand, In Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural System, pp. 45-52, 12-15 Sept., Ishikawa, Japan (peer reviewed).

Fukuda K, Taylor HH (2005) Phytoremediation of Land Contaminated by Heavy metals: Copper Concentrations in Self-sown Tomato Plants from a Landfill in New Zealand. In Proc. of Intl. Conf. on Research Highlights and Vanguard Technology on Environmental Engineering in Agricultural System, pp. 53-59, 12-15 Sept., Ishikawa, Japan (peer reviewed).

Hudson IL, Fukuda K, Dalrymple M (2005) Climate-pollution impacts on SIDS, In Proc. of 16th Biennial Congress on Modelling and Simulation (MODSIM05), pp. 286-292, 12-15 Dec., Melbourne (peer reviewed).

Fukuda K, Pearson PA (2006) Investigation of Singular Spectrum Analysis and Machine Learning for Road Sign Location. In Extended Abstracts of 7th Intl. Assoc. for Pattern Recognition workshop on Document Analysis Systems (DAS 2006), pp 29-32, 13-15 Feb., Nelson (peer reviewed extended abstract).

Fukuda K, Pearson PA (2006) Comparing data mining and image segmentation approaches for classifying defoliation in aerial forest imagery. In Proc. of 3rd Biennial meeting of the International Environmental Modelling and Software Society (iEMSs 2006), pp. 0-6, 9-12 Jul., Burlington, Vermont, USA (peer reviewed).

Fukuda K, Takaoka T (2007) Analysis of Air Pollution (PM_{10}) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm, In Proc. of the 2007 ACM Symposium on Applied Computing (SAC'07), pp. 153-157, 11-15 Mar., Seoul, Korea (peer reviewed).

Fukuda K (2007) Noise Reduction Approach for Decision Tree Construction: A Case Study of Knowledge Discovery on Climate and Air Pollution. In Proc. of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), pp. 697-704, 1-5 Apr., Hawaii (peer reviewed).

Fukuda K, Brown J (2007) Investigation of The Weed Risk Assessment Model Using Data Mining, 9th International Conference on Ecology and Management of Alien Plant Invasions (EMAPi9), 17-21 Sep, Perth (abstract).

Fukuda K, Takaoka T (2007) Investigation of the Maximum Association for Suicide Rate and Social Factors Using Computer Algorithm, International Congress on Modelling and Simulation (Modsim07), pp. 1381-1387, 10-13 Dec., Christchurch (peer reviewed).

Fukuda K, Brown J (2007) Classification Rule Extraction by Ant-Miner for Weed Risk Assessment, International Congress on Modelling and Simulation (Modsim07), pp. 2882-2888, 10-13 Dec., Christchurch (peer reviewed).

Fukuda K, Brown J, Williams P, Kean J (2008) The *K*-Maximum Subarray algorithm as an alternative clustering analysis for the spatial weed aggregation pattern, NZSA 2008 (abstract).

Fukuda K, Martin B (in press) Decision Trees as Information Source for Attribute Selection, In Proc. of the 2009 IEEE CIDM, 0-8.